FACULTAD DE ESTUDIOS ESTADÍSTICOS
MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2019/2020

Trabajo de Fin de Máster

TITULO: ANÁLISIS DE PREDICCIÓN DE CHURN EN UNA EMPRESA DE TELECOMUNICACIONES

***Alumna*: Michel Andrea Rodríguez Garzón**

***Tutor*: Antonio Sarasa**

Julio de 2020

UNIVERSIDAD COMPLUTENSE
MADRID

Contenido

Índice de Figuras

Índice de tablas

1. Introducción.

Las empresas con modelos de negocio basados en suscripción tienen el reto de mantener su base de clientes actual y captar nuevos, lo que lo convierte en un desafío grande teniendo en cuenta que los clientes cada vez son más exigentes y tienen mayor acceso a la información en cuanto a sus derechos como usuarios, funcionamiento del servicio, atención de quejas o reclamaciones, entidades a las cuales pueden acudir en segunda instancia cuando la reclamación interpuesta ante la compañía sea negativa, medios de contacto con la empresa, innovación de la empresa o de empresas competidoras, costos en el mercado de productos sustitutos y el voz a voz de familiares y conocidos que cuentan con un servicio similar.

De acuerdo a lo anterior es necesario contar con KPIS que permitan medir, controlar y tomar acciones preventivas o reactivas frente al resultado de la medición de los indicadores, uno de los más importantes es el Churn rate o tasa de cancelación el cual consiste en el porcentaje de clientes que se retira de la compañía en un periodo de tiempo determinado, este indicador suele ser medido de forma mensual y una de las formas de calcularlo consiste en dividir el número de clientes perdidos en un mes sobre el total de clientes a inicio de mes y este resultado multiplicarlo por el 100 por ciento, es decir si el total de clientes que abandonaron la compañía en el mes fueron 10 y la base de clientes a inicio de mes era 200 el Churn rate de este periodo fue 5%.

Podría decirse que este indicador permite medir la salud de la empresa, teniendo una base de clientes que cuenta con un servicio por suscripción se hace inevitable tener una rotación, sin embargo, el tener una tasa de cancelación baja mensual no implica que no tenga un impacto a largo plazo ya que si mensualmente se tiene un Churn de 1,2% en un año se tendrá 14,4%, lo cual afecta el rendimiento de la empresa.

Teniendo presente que el Churn depende del tipo de empresa en este trabajo nos enfocaremos en el sector de telecomunicaciones, es importante aclarar que las empresas suelen diferenciar la tasa de cancelación en voluntaria e involuntaria, la primera consiste en la decisión del suscriptor en cambiarse a otra empresa o proveedor de servicios, la segunda se genera cuando el suscriptor se traslada a un lugar en el cual no se tiene cobertura, muerte u otras, para este ejercicio nos concentraremos en el voluntario.

Dado que es mucho más costoso captar un nuevo cliente que retener uno existente las compañías invierten en campañas de retención que les permitan persuadir al cliente sobre su decisión, sin embargo se ven en la necesidad de contar con herramientas predictivas que les permitan identificar los patrones de comportamiento de los posibles clientes que quieren cancelar el servicio, esto les permitirá articular estrategias proactivas y mejorar su relación con el cliente, nos solo desde el enfoque de fidelización y retención, también áreas como ventas y marketing se pueden ver beneficiadas de dicho análisis ya que un cliente que tiene una probabilidad alta de abandono no sería bueno contactarlo para realizarle un ofrecimiento comercial.

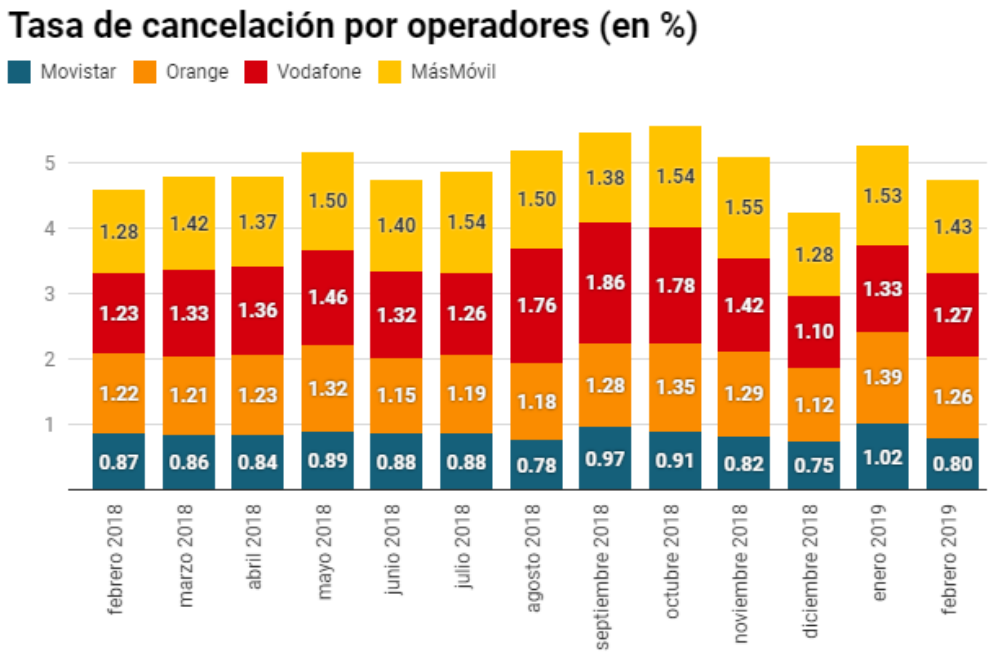


Figura 1 Churn empresas de telefonía móvil en España (inside economía digital)

Como vemos en la figura 1 [1] la empresa Movistar tiene un Churn bajo respectos a sus competidores lo que la llevo tener un 11,3% de tasa de cancelación anual si tomamos los periodos de febrero 2018 a febrero 2019, si realizamos el mismo ejercicio con la empresa Masmóvil estos tendrían un Churn de 18,6% lo cual supone una tasa muy alta de cancelación,
de allí la importancia de estimar modelos enfocados en la identificación de patrones que permitan anticiparse a la cancelación del cliente.

1.1. Estado del arte

El Machine learning ha sido fundamental y ampliamente utilizado para tratar los temas relacionados con la tasa de cancelación en las compañías de telecomunicaciones, donde la mayoría de las investigaciones se han orientado sobre aspectos demográficos, servicios contratados, tipo de contrato, medios de pago, incidencias tenidas en la compañía, como fallas en el servicio o problemas de facturación, etc.

Los modelos estadísticos tienen el objetivo de poder predecir y clasificar los posibles clientes que solicitaran la cancelación del servicio, dando paso a la identificación y mejora de procesos que están generando impacto negativo en la experiencia del cliente, mitigando estos se lograría una reducción del Churn y un fortalecimiento en la relación con el cliente.

Entre las investigaciones una se enfocaba en la metodología KDD (knowledge Discovery in Database), el cual es un proceso metodológico para encontrar un “modelo” válido, útil y entendible que describa patrones de acuerdo a la información, y como modelo entendemos que es la representación que intenta explicar ese patrón en los datos [2],

el principal desafío fue la unificación de las diferentes fuentes de datos, llegando a la conclusión que usar bases de datos operacionales instantáneas (no sujetas a cambios) son más útiles que usar bases de datos incrementales cuando se trata de datos transaccionales[3].

Por otro lado, una investigación tiene como objetivo profundizar en el análisis de la retención de clientes en servicios, por la trascendencia de sus implicaciones en la rentabilidad e imagen de las organizaciones, abordándolo de una forma cualitativa, analizando el abandono como un proceso, donde por medio de entrevistas a diferentes directivos de empresas y un estudio Delphi a mediadores le permite concluir que en general, los directivos de las empresas tienen un conocimiento poco profundo de los motivos relativos a la pérdida de clientes y se constata, además, que en las empresas, no hay modelos predictivos u otras técnicas o estudios sobre el abandono. En algunos casos, hay sistemas de alerta cuando el cliente tarda en pagar su subscrición y se solicita el pago del mismo [4], es de aclarar que es este trabajo no se abordara una técnica cualitativa sin embargo resulta de interés conocer otras técnicas para abordar esta problemática, así mismo sugiere la importancia para las empresas en contar con modelos predictivos que le permitan gestionar su tasa de cancelación y la relación con los clientes.

Por su parte, otra investigación realizada con la metodología CRISP-DM (Cross Industry Standard Process for Data Mining la cual consta de 6 fases, teniendo como fase primaria el entendimiento del negocio, en su mayoría las fases son bidireccionales y permiten revisar fases anteriores parcial o totalmente). El autor desea realizar la predicción del Churn en una empresa de salud aplicando los algoritmos KNN, un árbol simple, random forest, usando métodos de validación como el análisis de la curva ROC, índice Kappa y la matriz de confusión, en el cual concluye que el modelo KNN no funcionó en esa investigación, por lo que no se recomienda utilizarlo en modelos predictivos del churn con datos nominales. Por el contrario, el modelo predictivo realizado con el árbol de decisión simple con la heurística “Accuracy”, completó con éxito todos los criterios de evaluación planteados, teniendo un nivel de certeza del 97.52% [5].

Al otro lado, realizaron una investigación en la cual se quiso comparar el algoritmo adaboost desbalanceado y la regresión logística asimétrica e identificar cuál de ellas tiene mayor precisión en la predicción de fuga de clientes en el sector de telefonía móvil, el autor usa métodos de muestreo para el balanceo de las muestras para el primer algoritmo y en el segundo ajusta y/o modifica el algoritmo o función, concluyendo que el algoritmo adaboost que tiene como objetivo por medio del entrenamiento iterativo de los clasificadores débiles o de base, asignar mayor importancia a los datos mal clasificados de esta forma obteniendo un nuevo clasificador, el algoritmo es superior y más eficiente que la regresión logística para la clasificación de clientes que se fugan en la empresa de telefonía [6].

Analizando la investigación sobre la aplicación de algoritmos de ML sobre el churn en los planes de ahorro en un concesionario, estos enfatizan en la problemática de estos planes de ahorro donde se identifica que antes de los 12 meses se tiene la mayor probabilidad de cancelar, para llegar a esta conclusión trabajaron con árboles de

decisión, naive vayes y redes neuronales, siendo el primer algoritmo el que mejores resultados obtuvo con una exactitud de clasificación del 99,08% y una tasa de error del 0,2 y les permitió la identificación de patrones de clientes que desean cancelar el ahorro, por otro lado las redes neuronales tuvieron un buen desempeño sin embargo los autores indican que son cajas negras que no permiten ver como las variables de entrada afectan a los resultados del modelo[7].

1.2. Objetivos

El objetivo principal de este proyecto es predecir los posibles clientes que podrían cancelar la suscripción en una empresa de telecomunicaciones, para conseguir el objetivo se tendrán como objetivos secundarios: obtención de los datos, la preparación y análisis exploratorio de las variables input, aplicación de algoritmos de predicción como regresión logística, redes neuronales, random forest, gradient boosting, entre otros , finalizando en la evaluación y comparación de los distintos modelos.

2. Metodología

Con el fin de alcanzar los objetivos propuestos se trabaja con la metodología SEMMA, esta fue desarrollada por el instituto SAS donde el acrónimo traduce Sample, Explorer, Modify, Model, Asses, que al pasarlo al castellano se interpreta como Muestra, Explora, Modifica, Modela y Evalúa, cabe aclara que es un proceso flexible en el cual se puede regresar a la fase anterior para modificarla.

- Sample: Esta etapa consiste en preparar los datos para su exploración, teniendo presente el volumen de los datos ya que en caso de ser muy grandes se debe tomar una muestra representativa, en este caso contamos con una muestra adecuada para el proyecto.

- Explorer: Corresponde a la exploración grafica de los datos, cuya finalidad en la identificación de relaciones y anomalías en las variables, esta etapa fue realizada en SAS Enterprise Miner.

- Modify: Modificación de los datos para el tratamiento de valores atípicos o faltantes, al igual que la transformación y/o creación de variables que faciliten la modelización de los datos, etapa realizada en SAS Enterprise Miner.

- Model: Generación de modelos que permitan predecir la variable objetivo que para este caso es el Churn, los modelos fueron realizados en SAS y R.

- Asses: Se comprueba la calidad de las predicciones y se comparan los modelos por medio de estadísticos como AUC, la tasa de fallos entre otros, realizando previamente validación cruzada repetida.

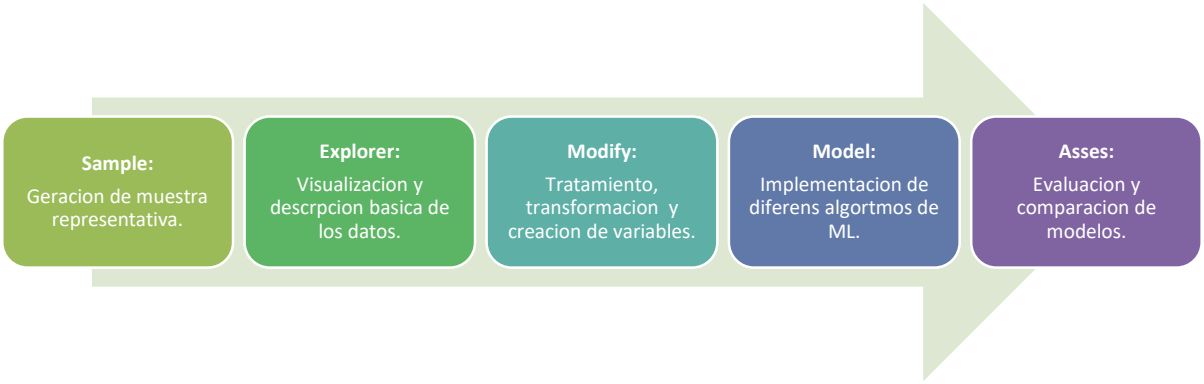


Figura 2 Esquema de metodología SEMMA

2.1. Algoritmos de Machine Learning

Regresión logística:

La Regresión Logística es un método estadístico para predecir clases binarias. El resultado o variable objetivo es de naturaleza dicotómica, donde describe y estima la relación entre una variable binaria dependiente y las variables independientes.

Por su parte la Regresión Logística lleva el nombre de la función utilizada en el núcleo del método, la función logística es también llamada función Sigmoide. Esta función es una curva en forma de S que puede tomar cualquier número de valor real y asignar a un valor entre 0 y 1.

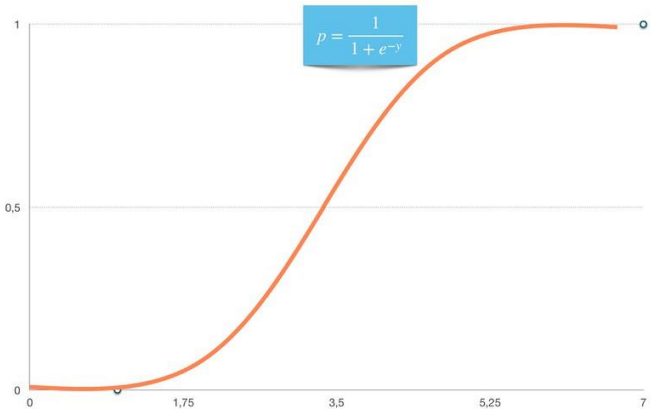


Figura 3 Ecuación de la función del sigmoide [8]

Si la curva va a infinito positivo la predicción se convertirá en 1, y si la curva pasa el infinito negativo, la predicción se convertirá en 0. Si la salida de la función Sigmoide es mayor que 0.5, podemos clasificar el resultado como 1 o SI, y si es menor que 0.5 podemos clasificarlo como 0 o NO [8].

Redes neuronales:

Paradigma de aprendizaje y procesamiento automático inspirado en el funcionamiento del sistema nervioso humano, una red neuronal está compuesta por un conjunto de neuronas interconectadas entre sí mediante enlaces, cada neurona toma como entradas las salidas de las neuronas de las capas antecesoras, cada una de esas entradas se multiplica por un peso, se agregan los resultados parciales y mediante una función de activación se calcula la salida. Esta salida a su vez es entrada de la neurona a la que precede, La unión de todas estas neuronas interconectadas es lo que compone la red neuronal artificial, las redes neuronales no son más que redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico [9].

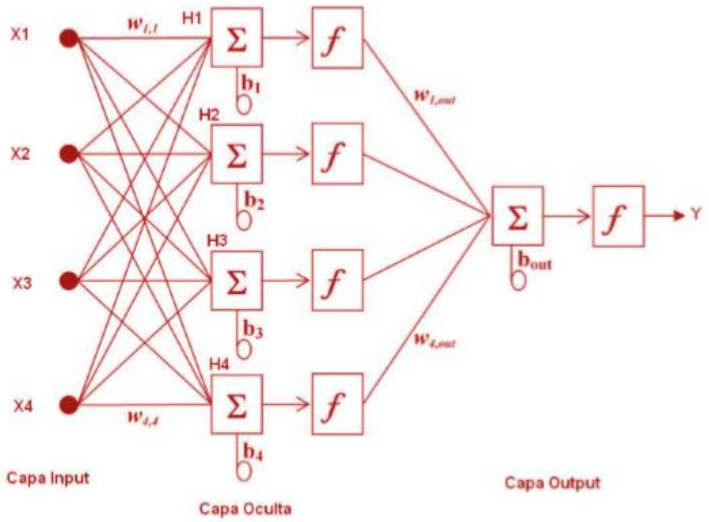


Figura 4 Esquema red neuronal (Portela 2019)

Conceptos básicos de las redes:

- Conjunto de entradas (x): Representan las entradas de la red neuronal.
- Pesos sinápticos (w): Cada entrada tiene un peso que se va ajustando de forma automática a medida que la red neuronal va aprendiendo.
- Función de agregación (Σ): Realiza el sumatorio de todas las entradas ponderadas por sus pesos.
- Función de activación (F): Se encarga de mantener el conjunto de valores de salida en un rango determinado, normalmente (0,1) o (-1,1).
- Salida (Y): Representa el valor resultante tras pasar por la red neuronal.

Bagging:

Modelo predictivo creado por Leo Breiman que denominó Bagging (o Bootstrap Aggregating). Esta técnica consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.

La técnica de Bagging sigue estos pasos:

A. Divide el set de Entrenamiento en distintos sub set de datos, obteniendo como resultado diferentes muestras aleatorias con las siguientes características:
- Muestra uniforme (misma cantidad de individuos en cada set).
- Muestras con reemplazo (los individuos pueden repetirse en el mismo set de datos).
- El tamaño de la muestra es igual al tamaño del set de entrenamiento, pero no contiene a todos los individuos ya que algunos se repiten.
- Si se usan muestras sin reemplazo, suele elegirse el 50% de los datos como tamaño de muestra.

B. Luego se crea un modelo predictivo con cada set, obteniendo modelos diferentes.
C. Luego se construye o ensambla un único modelo predictivo, que es el promedio de todos los modelos [10].

Random Forest:

Es un modelo basado en arboles cuya diferencia con bagging consiste en incorporar aleatoriedad en las variables utilizadas para segmentar cada árbol, funcionado de la siguiente forma:

Dados los datos de tamaño N,

A. Repetir m veces i), ii), iii):
   (i) Seleccionar N observaciones con reemplazamiento de los datos originales.
   (ii) Aplicar un árbol de la siguiente manera: En cada nodo, seleccionar p variables de las k originales y de las p elegidas, escoger la mejor variable para la partición del nodo.
   (iii) Obtener predicciones para todas las observaciones originales N.
B. Promediar las m predicciones obtenidas en el apartado A) [11].

Gradient Boosting:

El algoritmo gradient boosting consiste en repetir la construcción de árboles de regresión/clasificación, modificando ligeramente las predicciones iniciales cada vez, intentando ir minimizando los residuos en la dirección de decrecimiento.

Al plantear diferentes árboles cada vez, el proceso va ajustando las predicciones cada vez más a los datos, y de alguna manera unos árboles corrigen a otros con lo cual la flexibilidad y adaptación del método mejora respecto a la construcción de un único árbol.

Los principales parámetros a controlar son la constante de regularización v(shrink), el número de iteraciones y las características propias de los arboles como el número de

hojas finales, el número de divisiones máxima por nodo, el p-valor para las divisiones de cada nodo y el número de observaciones mínimo de cada rama. [11].

Sopor vector machine:

Una máquina de vectores de soporte (SVM) es un algoritmo de aprendizaje supervisado que se puede emplear para clasificación binaria o regresión, la cual construye un hiperplano óptimo en forma de superficie de decisión, de modo que el margen de separación entre las dos clases en los datos se amplía al máximo. Los vectores de soporte hacen referencia a un pequeño subconjunto de las observaciones de entrenamiento que se utilizan como soporte para la ubicación óptima de la superficie de decisión.

El entrenamiento de una máquina de vectores de soporte consta de dos fases:

1. Transformar los predictores (datos de entrada) en un espacio de características altamente dimensional. En esta fase es suficiente con especificar el kernel; los datos nunca se transforman explícitamente al espacio de características. Este proceso se conoce comúnmente como el truco kernel.
2. Resolver un problema de optimización cuadrática que se ajuste a un hiperplano óptimo para clasificar las características transformadas en dos clases. El número de características transformadas está determinado por el número de vectores de soporte.

Para construir la superficie de decisión solo se requieren los vectores de soporte seleccionados de los datos de entrenamiento.

Entre los kernels populares que se emplean en este trabajo son [12]:

Tipo de SVM	Kernel Mercer	Descripción
Función de base radial (RBF) o gaussiana	$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$	Aprendizaje de una sola clase; σes la anchura del kernel
Lineal	$K(x_1, x_2) = x_1^{\mathsf{T}} x_2$	Aprendizaje de dos clases
Polinómica	$K(x_1, x_2) = \left(x_1^{\mathsf{T}} x_2 + 1\right)^{\rho}$	ρ es el orden del polinomio

Figura 5 Tipos de Kernel en SVM

Ensamblado:

los métodos de ensamblado utilizan múltiples algoritmos de aprendizaje para obtener un rendimiento predictivo que mejore el que podría obtenerse por medio de cualquiera de los algoritmos de aprendizaje individuales que lo constituyen. [13]

Existen técnicas básicas de combinado de modelos como Bagging, Boosting, Stacking y otros, en este trabajo no centraremos en Stacking, el cual consiste en la combinación de las predicciones de varios algoritmos, existiendo tres ideas básicas:

1) Averaging (promediado): Se calcula el promedio de las predicciones, Si se trata de clasificación, se obtiene el promedio de las probabilidades. Se puede utilizar también promedio ponderado.
2) Voto (para clasificación): Se predice el resultado con mayoría entre las predicciones: y1=0, y2=0, y3=1 predicción=1.
3) Combinación a partir de otro algoritmo (esto es estrictamente stacking) [11].

3. Preparación de los datos

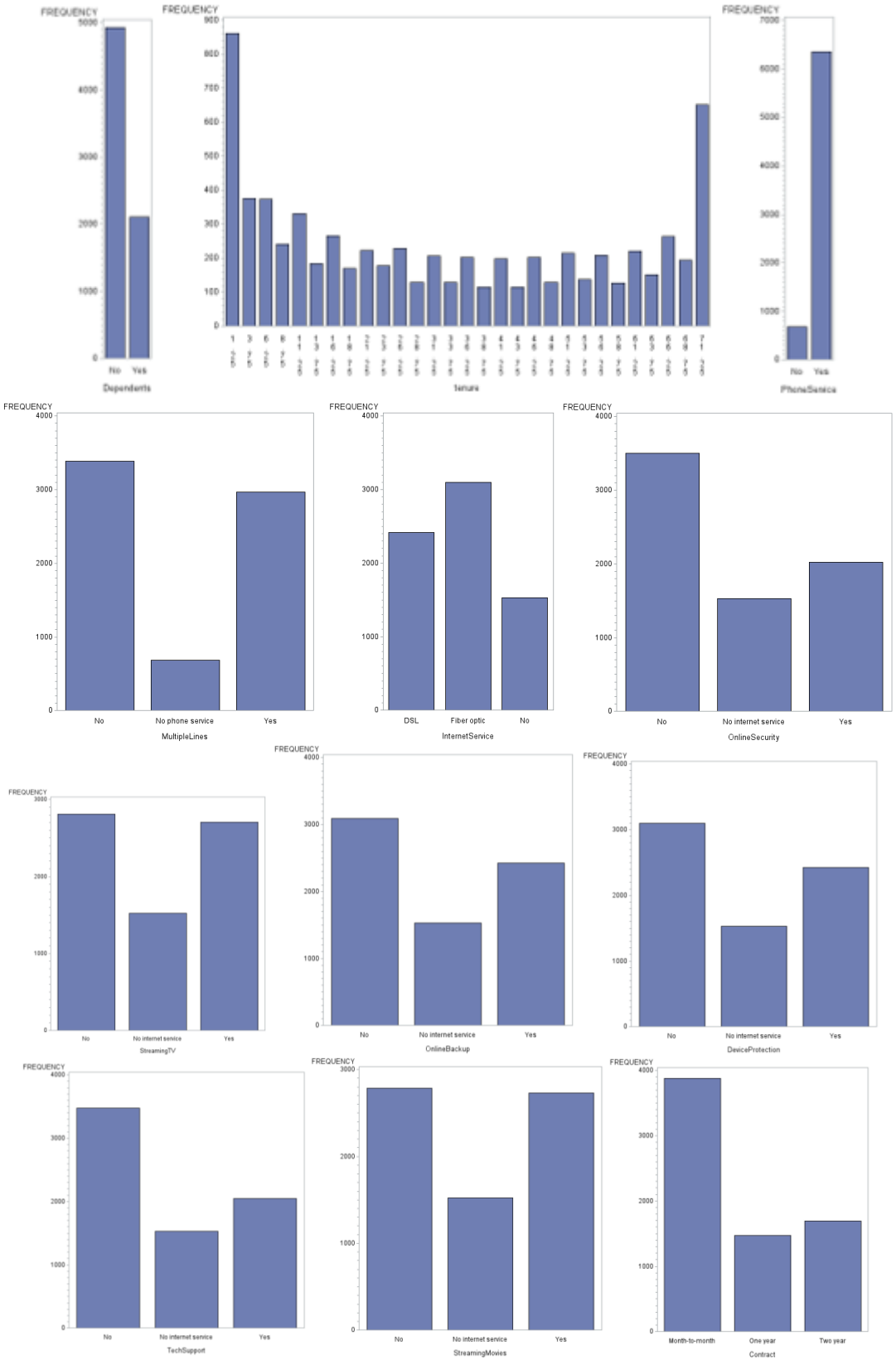
3.1. Origen de los datos.

Los datos utilizados en ente proyecto provienen de la comunidad en línea de científicos de datos Kaggle, esta base de datos consta de 7043 registros y 21 variables incluyendo la variable objetivo contienen información sobre los servicios contratados por el cliente, información de la cuenta e información demográfica, originalmente las variables se encuentran en inglés, pero fueron traducidas al castellano para la descripción las cuales relaciono a continuación:

Variable	Descripción	Tipo
Identificación del cliente	ID	
Variable objetivo		
Churn	Clientes que se retiraron de la empresa, variable objetivo.	Clase
Informacion de la cuenta del cliente		
Antigüedad	Número de meses que el cliente ha permanecido en la empresa.	Intervalo
Contrato	Plazo del contrato del cliente (mes a mes, un año, dos años).	Clase
Método de pago	Método del pago del cliente (cheque, transferencia bancaria, tarjeta de crédito).	Clase
Facturación sin papel	Si el cliente tiene facturación electrónica.	Clase
Cargos mensuales	Monto cobrado al cliente mensualmente.	Intervalo
Cargos totales	El imposte total cobrado al cliente.	Intervalo
Informacion demografica		
Género	Si el cliente es hombre o mujer.	Clase
Ciudadano mayor	Si el cliente es una persona mayor o no.	Clase
Compañero	Si el cliente tiene un socio o no. (lo entiendo es si tiene pareja).	Clase
Dependientes	Si el cliente tiene dependientes (lo entiendo si el cliente tiene hijos)	Clase
Servicios suscritos por el cliente		
Servicio telefónico	Si el cliente tiene un servicio telefónico (si/no).	Clase
Múltiples líneas	Si el cliente tiene varias líneas (si/no).	Clase
Servicio de internet	Proveedor de servicios de internet del cliente (dsl, fibra óptica, no).	Clase
Seguridad en línea	Si el cliente tiene seguridad en línea o no.	Clase
Onlinebackup	Si el cliente tiene un respaldo en línea.	Clase
Protección del dispositivo	Si el cliente tiene protección del dispositivo o no.	Clase
Apoyo técnico	Si el cliente tiene apoyo técnico (si, no, sin servicio de internet).	Clase
StreamingTV	Si el cliente tiene transmisión de tv (si, no, sin servicio de internet).	Clase
Streaming películas	Si el cliente tiene películas en streaming (si, no, sin servicio de internet).	Clase

Tabla 1 Descripción de variables

Se realizan los histogramas de las variables de clase con el fin de tener una idea previa al análisis exploratorio la calidad de la información en términos de niveles adicionales a los anteriormente relacionados.


FREQUENCY
Dependents
No
Yes
Tenure
PhoneService
MultipleLines
No phone service
InternetService
DSL
Fiber optic
OnlineSecurity
No internet service
StreamingTV
OnlineBackup
DeviceProtection
TechSupport
StreamingMovies
Contract
Month-to-month
One year
Two year

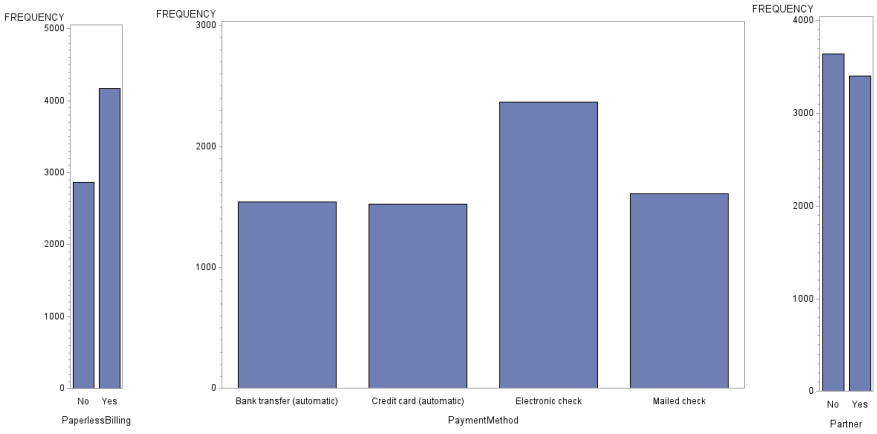


Figura 6 Histogramas de variables de clase

En los histogramas anteriores podemos ver no se encuentran niveles adicionales en las variables.

3.2. Análisis exploratorio.

Para este análisis los datos son cargados a SAS Enterprise Miner para realizar su depuración.

3.2.1. Análisis descriptivo del conjunto de datos

Haciendo uso del nodo DMDB se detectan los missings de las variables y los posibles errores en ellas, el 5% de las 7043 observaciones del dataset son 352 lo que nos permite observar que ninguna variable tiene esa cantidad de missings estando entonces dentro del rango ideal. Respecto a las variables de intervalo vemos que la asimetría se mantiene dentro del rango de -1 y 1 y la variable TotalChargues presenta 11 ausentes.

Variables de intervalo

Variable	Etiqueta	Ausente	N	Mínimo	Máximo ▼	Media	Desviación estándar	Asimetría	Curtosis
TotalCharges	TotalCharges	11	7032	19	8685	2283.321	2266.769	0.961652	-0.23177
MonthlyCharg...	MonthlyCha...	0	7043	18	119	64.78305	30.0943	-0.22033	-1.25688
tenure	tenure	0	7043	0	72	32.37115	24.55948	0.23954	-1.38737

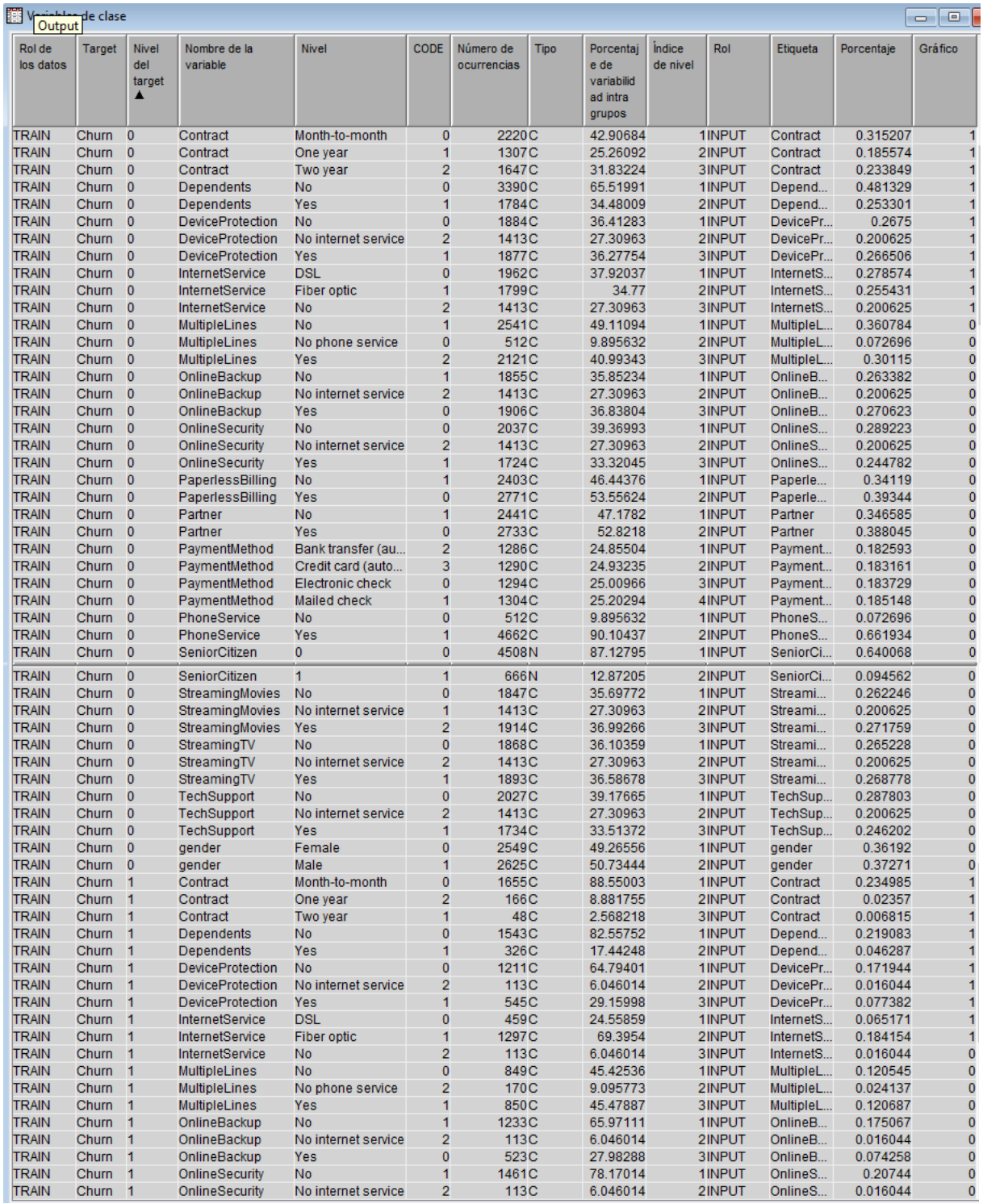
Tabla 2 Estadísticos de variables de intervalo

Variables de clase

Variable	Etiqueta	Tipo	Número de niveles	Ausente
Churn	Churn	N	2	0
Contract	Contract	C	3	0
Dependents	Dependents	C	2	0
DeviceProtection	DeviceProtection	C	3	0
InternetService	InternetService	C	3	0
MultipleLines	MultipleLines	C	3	0
OnlineBackup	OnlineBackup	C	3	0
OnlineSecurity	OnlineSecurity	C	3	0
PaperlessBilling	PaperlessBilling	C	2	0
Partner	Partner	C	2	0
PaymentMethod	PaymentMethod	C	4	0
PhoneService	PhoneService	C	2	0
SeniorCitizen	SeniorCitizen	N	2	0
StreamingMovies	StreamingMovies	C	3	0
StreamingTV	StreamingTV	C	3	0
TechSupport	TechSupport	C	3	0
gender	gender	C	2	0

Tabla 3 Estadísticos de variables de clase

El nodo explorador de estadísticos nos da más información de las variables de clase, al revisarlas se confirma que no hay errores para corregir.


Variables de clase

Output

Rol de los datos	Target	Nivel del target ▲	Nombre de la variable	Nivel	CODE	Número de ocurrencias	Tipo	Porcentaje de variabilidad intra grupos	Índice de nivel	Rol	Etiqueta	Porcentaje	Gráfico
TRAIN	Churn	0	Contract	Month-to-month	0	2220	C	42.90684	1	INPUT	Contract	0.315207	1
TRAIN	Churn	0	Contract	One year	1	1307	C	25.26092	2	INPUT	Contract	0.185574	1
TRAIN	Churn	0	Contract	Two year	2	1647	C	31.83224	3	INPUT	Contract	0.233849	1
TRAIN	Churn	0	Dependents	No	0	3390	C	65.51991	1	INPUT	Depend...	0.481329	1
TRAIN	Churn	0	Dependents	Yes	1	1784	C	34.48009	2	INPUT	Depend...	0.253301	1
TRAIN	Churn	0	DeviceProtection	No	0	1884	C	36.41283	1	INPUT	DevicePr...	0.2675	1
TRAIN	Churn	0	DeviceProtection	No internet service	2	1413	C	27.30963	2	INPUT	DevicePr...	0.200625	1
TRAIN	Churn	0	DeviceProtection	Yes	1	1877	C	36.27754	3	INPUT	DevicePr...	0.266506	1
TRAIN	Churn	0	InternetService	DSL	0	1962	C	37.92037	1	INPUT	InternetS...	0.278574	1
TRAIN	Churn	0	InternetService	Fiber optic	1	1799	C	34.77	2	INPUT	InternetS...	0.255431	1
TRAIN	Churn	0	InternetService	No	2	1413	C	27.30963	3	INPUT	InternetS...	0.200625	1
TRAIN	Churn	0	MultipleLines	No	1	2541	C	49.11094	1	INPUT	MultipleL...	0.360784	0
TRAIN	Churn	0	MultipleLines	No phone service	0	512	C	9.895632	2	INPUT	MultipleL...	0.072696	0
TRAIN	Churn	0	MultipleLines	Yes	2	2121	C	40.99343	3	INPUT	MultipleL...	0.30115	0
TRAIN	Churn	0	OnlineBackup	No	1	1855	C	35.85234	1	INPUT	OnlineB...	0.263382	0
TRAIN	Churn	0	OnlineBackup	No internet service	2	1413	C	27.30963	2	INPUT	OnlineB...	0.200625	0
TRAIN	Churn	0	OnlineBackup	Yes	0	1906	C	36.83804	3	INPUT	OnlineB...	0.270623	0
TRAIN	Churn	0	OnlineSecurity	No	0	2037	C	39.36993	1	INPUT	OnlineS...	0.289223	0
TRAIN	Churn	0	OnlineSecurity	No internet service	2	1413	C	27.30963	2	INPUT	OnlineS...	0.200625	0
TRAIN	Churn	0	OnlineSecurity	Yes	1	1724	C	33.32045	3	INPUT	OnlineS...	0.244782	0
TRAIN	Churn	0	PaperlessBilling	No	1	2403	C	46.44376	1	INPUT	Paperle...	0.34119	0
TRAIN	Churn	0	PaperlessBilling	Yes	0	2771	C	53.55624	2	INPUT	Paperle...	0.39344	0
TRAIN	Churn	0	Partner	No	1	2441	C	47.1782	1	INPUT	Partner	0.346585	0
TRAIN	Churn	0	Partner	Yes	0	2733	C	52.8218	2	INPUT	Partner	0.388045	0
TRAIN	Churn	0	PaymentMethod	Bank transfer (au...	2	1286	C	24.85504	1	INPUT	Payment...	0.182593	0
TRAIN	Churn	0	PaymentMethod	Credit card (auto...	3	1290	C	24.93235	2	INPUT	Payment...	0.183161	0
TRAIN	Churn	0	PaymentMethod	Electronic check	0	1294	C	25.00966	3	INPUT	Payment...	0.183729	0
TRAIN	Churn	0	PaymentMethod	Mailed check	1	1304	C	25.20294	4	INPUT	Payment...	0.185148	0
TRAIN	Churn	0	PhoneService	No	0	512	C	9.895632	1	INPUT	PhoneS...	0.072696	0
TRAIN	Churn	0	PhoneService	Yes	1	4662	C	90.10437	2	INPUT	PhoneS...	0.661934	0
TRAIN	Churn	0	SeniorCitizen	0	0	4508	N	87.12795	1	INPUT	SeniorCi...	0.640068	0
TRAIN	Churn	0	SeniorCitizen	1	1	666	N	12.87205	2	INPUT	SeniorCi...	0.094562	0
TRAIN	Churn	0	StreamingMovies	No	0	1847	C	35.69772	1	INPUT	Streami...	0.262246	0
TRAIN	Churn	0	StreamingMovies	No internet service	1	1413	C	27.30963	2	INPUT	Streami...	0.200625	0
TRAIN	Churn	0	StreamingMovies	Yes	2	1914	C	36.99266	3	INPUT	Streami...	0.271759	0
TRAIN	Churn	0	StreamingTV	No	0	1868	C	36.10359	1	INPUT	Streami...	0.265228	0
TRAIN	Churn	0	StreamingTV	No internet service	2	1413	C	27.30963	2	INPUT	Streami...	0.200625	0
TRAIN	Churn	0	StreamingTV	Yes	1	1893	C	36.58678	3	INPUT	Streami...	0.268778	0
TRAIN	Churn	0	TechSupport	No	0	2027	C	39.17665	1	INPUT	TechSup...	0.287803	0
TRAIN	Churn	0	TechSupport	No internet service	2	1413	C	27.30963	2	INPUT	TechSup...	0.200625	0
TRAIN	Churn	0	TechSupport	Yes	1	1734	C	33.51372	3	INPUT	TechSup...	0.246202	0
TRAIN	Churn	0	gender	Female	0	2549	C	49.26556	1	INPUT	gender	0.36192	0
TRAIN	Churn	0	gender	Male	1	2625	C	50.73444	2	INPUT	gender	0.37271	0
TRAIN	Churn	1	Contract	Month-to-month	0	1655	C	88.55003	1	INPUT	Contract	0.234985	1
TRAIN	Churn	1	Contract	One year	2	166	C	8.881755	2	INPUT	Contract	0.02357	1
TRAIN	Churn	1	Contract	Two year	1	48	C	2.568218	3	INPUT	Contract	0.006815	1
TRAIN	Churn	1	Dependents	No	0	1543	C	82.55752	1	INPUT	Depend...	0.219083	1
TRAIN	Churn	1	Dependents	Yes	1	326	C	17.44248	2	INPUT	Depend...	0.046287	1
TRAIN	Churn	1	DeviceProtection	No	0	1211	C	64.79401	1	INPUT	DevicePr...	0.171944	1
TRAIN	Churn	1	DeviceProtection	No internet service	2	113	C	6.046014	2	INPUT	DevicePr...	0.016044	1
TRAIN	Churn	1	DeviceProtection	Yes	1	545	C	29.15998	3	INPUT	DevicePr...	0.077382	1
TRAIN	Churn	1	InternetService	DSL	0	459	C	24.55859	1	INPUT	InternetS...	0.065171	1
TRAIN	Churn	1	InternetService	Fiber optic	1	1297	C	69.3954	2	INPUT	InternetS...	0.184154	1
TRAIN	Churn	1	InternetService	No	2	113	C	6.046014	3	INPUT	InternetS...	0.016044	0
TRAIN	Churn	1	MultipleLines	No	0	849	C	45.42536	1	INPUT	MultipleL...	0.120545	0
TRAIN	Churn	1	MultipleLines	No phone service	2	170	C	9.095773	2	INPUT	MultipleL...	0.024137	0
TRAIN	Churn	1	MultipleLines	Yes	1	850	C	45.47887	3	INPUT	MultipleL...	0.120687	0
TRAIN	Churn	1	OnlineBackup	No	1	1233	C	65.97111	1	INPUT	OnlineB...	0.175067	0
TRAIN	Churn	1	OnlineBackup	No internet service	2	113	C	6.046014	2	INPUT	OnlineB...	0.016044	0
TRAIN	Churn	1	OnlineBackup	Yes	0	523	C	27.98288	3	INPUT	OnlineB...	0.074258	0
TRAIN	Churn	1	OnlineSecurity	No	1	1461	C	78.17014	1	INPUT	OnlineS...	0.20744	0
TRAIN	Churn	1	OnlineSecurity	No internet service	2	113	C	6.046014	2	INPUT	OnlineS...	0.016044	0

TRAIN	Churn	1	OnlineSecurity	Yes	0	295C	15.78384	3INPUT	OnlineS...	0.041886	0
TRAIN	Churn	1	PaperlessBilling	No	1	469C	25.09363	1INPUT	Paperle...	0.066591	0
TRAIN	Churn	1	PaperlessBilling	Yes	0	1400C	74.90637	2INPUT	Paperle...	0.198779	0
TRAIN	Churn	1	Partner	No	0	1200C	64.20546	1INPUT	Partner	0.170382	0
TRAIN	Churn	1	Partner	Yes	1	669C	35.79454	2INPUT	Partner	0.094988	0
TRAIN	Churn	1	PaymentMethod	Bank transfer (au...	2	258C	13.80417	1INPUT	Payment...	0.036632	0
TRAIN	Churn	1	PaymentMethod	Credit card (auto...	3	232C	12.41306	2INPUT	Payment...	0.032941	0
TRAIN	Churn	1	PaymentMethod	Electronic check	1	1071C	57.30337	3INPUT	Payment...	0.152066	0
TRAIN	Churn	1	PaymentMethod	Mailed check	0	308C	16.4794	4INPUT	Payment...	0.043731	0
TRAIN	Churn	1	PhoneService	No	1	170C	9.095773	1INPUT	PhoneS...	0.024137	0
TRAIN	Churn	1	PhoneService	Yes	0	1699C	90.90423	2INPUT	PhoneS...	0.241232	0
TRAIN	Churn	1	SeniorCitizen	0	0	1393N	74.53184	1INPUT	SeniorCi...	0.197785	0
TRAIN	Churn	1	SeniorCitizen	1	1	476N	25.46816	2INPUT	SeniorCi...	0.067585	0
TRAIN	Churn	1	StreamingMovies	No	0	938C	50.18727	1INPUT	Streami...	0.133182	0
TRAIN	Churn	1	StreamingMovies	No internet service	2	113C	6.046014	2INPUT	Streami...	0.016044	0
TRAIN	Churn	1	StreamingMovies	Yes	1	818C	43.76672	3INPUT	Streami...	0.116144	0
TRAIN	Churn	1	StreamingTV	No	0	942C	50.40128	1INPUT	Streami...	0.13375	0
TRAIN	Churn	1	StreamingTV	No internet service	2	113C	6.046014	2INPUT	Streami...	0.016044	0
TRAIN	Churn	1	StreamingTV	Yes	1	814C	43.5527	3INPUT	Streami...	0.115576	0
TRAIN	Churn	1	TechSupport	No	0	1446C	77.36758	1INPUT	TechSup...	0.20531	0
TRAIN	Churn	1	TechSupport	No internet service	2	113C	6.046014	2INPUT	TechSup...	0.016044	0
TRAIN	Churn	1	TechSupport	Yes	1	310C	16.58641	3INPUT	TechSup...	0.044015	0
TRAIN	Churn	1	gender	Female	1	939C	50.24077	1INPUT	gender	0.133324	0
TRAIN	Churn	1	gender	Male	0	930C	49.75923	2INPUT	gender	0.132046	0

Tabla 4 Identificación de niveles de variables de clase

Tratamiento de datos atípicos y faltantes

Para el tratamiento de datos atípicos se utiliza el nodo reemplazo y la técnica desviación estándar debido a que los datos son simétricos, al realizar este proceso podemos evidenciar la cantidad de datos atípicos que han sido transformados a valores ausentes "missing", en este caso no tenemos missing como los vemos en el siguiente cuadro.

Cuentas de reemplazo total

Variable	Etiqueta	Rol	Entrenamiento
MonthlyCharges	MonthlyCharges	INPUT	0
TotalCharges	TotalCharges	INPUT	0
tenure	tenure	INPUT	0

Tabla 5 Datos atípicos variables de intervalo

Seguido he usado el nodo de código SAS para crear la variable numMissing y verifico la cantidad de ausentes de cada observación con el nodo DMBD la variable creada tiene un máximo de 1 los cual no supera el 5% de las variables.

Variables de intervalo

Variable	Etiqueta	Ausente	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis
REP_Month...	Replaceme...	0	7043	18	119	64.78305	30.0943	-0.22033	-1.25688
REP_Total...	Replaceme...	11	7032	19	8685	2283.321	2266.769	0.961652	-0.23177
REP_tenure	Replaceme...	0	7043	0	72	32.37115	24.55948	0.23954	-1.38737
numMissing		0	7043	0	1	0.001562	0.039492	25.24967	635.7264

Tabla 6 Cantidad de datos ausentes

Debido a que se tienen datos ausentes se hace uso del nodo imputar, el método de imputación usado es la distribución para las variables de intervalo, en las variables de clase no tenemos datos ausentes, no se genera la variable indicadora porque ninguna de las variables tiene más del 5% de las observaciones como missing.

Resumen de imputación

Nombre de la variable	Imputar método	Variable imputada	Número de ausentes para TRAIN	Variable imputada	Rol	Nivel de medida	Etiqueta
REP_TotalCharges	DISTRIBUTION	IMP_REP_TotalCharges	11		INPUT	INTERVAL	Replacem...

Tabla 7 Imputación de datos

Posterior a la imputación se hace uso del nodo DMBD para verificar que los datos han quedado limpios de atípicos y ausentes como poderlo a continuación.

Variables de intervalo

Variable	Etiqueta	Ausente	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis
IMP_REP_TotalCharges	Imputed: Replace...	0	7043	19	8685	2282.75	2266.163	0.961933	-0.23072
REP_MonthlyCharges	Replacement: Mo...	0	7043	18	119	64.78305	30.0943	-0.22033	-1.25688
REP_tenure	Replacement: ten...	0	7043	0	72	32.37115	24.55948	0.23954	-1.38737
numMissing		0	7043	0	1	0.001562	0.039492	25.24967	635.7264

Output

Variable	Etiqueta	Tipo	Número de niveles	Ausente
Churn	Churn	N	2	0
Contract	Contract	C	3	0
Dependents	Dependents	C	2	0
DeviceProtection	DeviceProtection	C	3	0
InternetService	InternetService	C	3	0
MultipleLines	MultipleLines	C	3	0
OnlineBackup	OnlineBackup	C	3	0
OnlineSecurity	OnlineSecurity	C	3	0
PaperlessBilling	PaperlessBilling	C	2	0
Partner	Partner	C	2	0
PaymentMethod	PaymentMethod	C	4	0
PhoneService	PhoneService	C	2	0
SeniorCitizen	SeniorCitizen	N	2	0
StreamingMovies	StreamingMovies	C	3	0
StreamingTV	StreamingTV	C	3	0
TechSupport	TechSupport	C	3	0
gender	gender	C	2	0

Tabla 8 Verificación de ausentes posterior a la imputación

Se crea un variable aleatoria que nos ayudara a identificar que variables serán importantes en el modelo, para esto se hace uso del nodo transformar variable con la fórmula de numero aleatorio, se trabaja con la semilla (12345) y numero de clases de agrupamiento optimo 5.

Estadísticos de transformaciones

Fuente	Método	Nombre de la variable	Fórmula	Número de niveles	No ausente	Ausente	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis	Etiqueta
Output	Formula	var_aleatoria	DMRAN(12345)	.	7043	0	.0000502	0.999986	0.499718	0.287752	0.0068	-1.18369	

Tabla 9 Generación de variable aleatoria

Con el nodo guardar datos se guardan los datos y el diagrama de depuración queda de la siguiente forma:

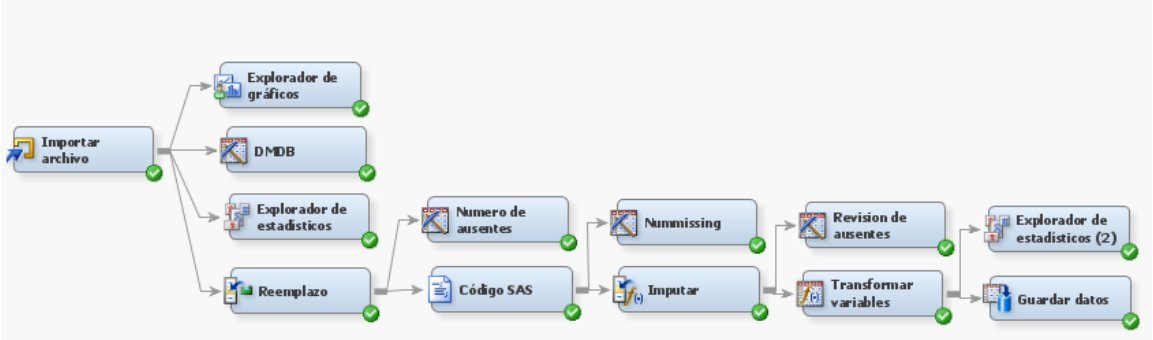


Diagrama 1 Depuración en Miner

3.2.2. Analisis de la relación entre las variables

Con el nodo explorador de estadísticos se quiere ver que variables podrían ser útiles para el modelo, se analiza el grafico chi-cuadrado que nos permite ver la relación entre las variables de clase e intervalo con la variable objetivo, entre ella vemos que las variables que tienen mayor dependencia son: contract, onlineSecurity, TechSupport, InternetService, PaymentMethod, Onlineback up, DeviceProtection, StreamingMovies, StreamingTV, las otras variables son más independientes de la variable objetivo y no le aportan mucho valor, en especial MultipleLines, PhoneServices, gerner.

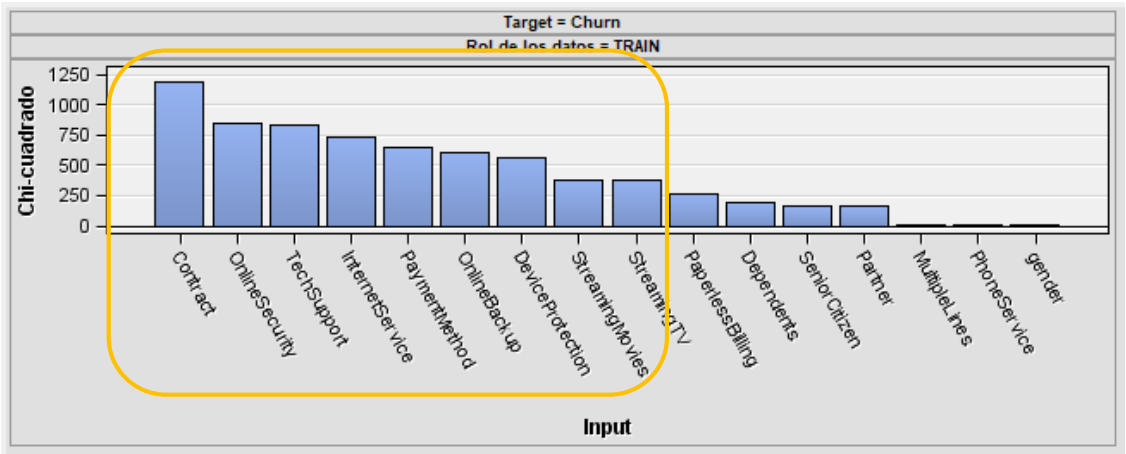


Figura 7 Grafico Chi-cuadrado

Con el grafico V de Cramer podemos ver el orden de importancia de las variables, a diferencia del grafico anterior se incluyen variables que tienen mayor relación con la variable objetivo como Rep_tenure, Rep_monthlyCharges, y IMP_REP_totalchargues, en este grafico aparecen las variables var_aleatoria y numMissing las cuales se encentran casi al final ya que no aportan información a la variable objetivo.

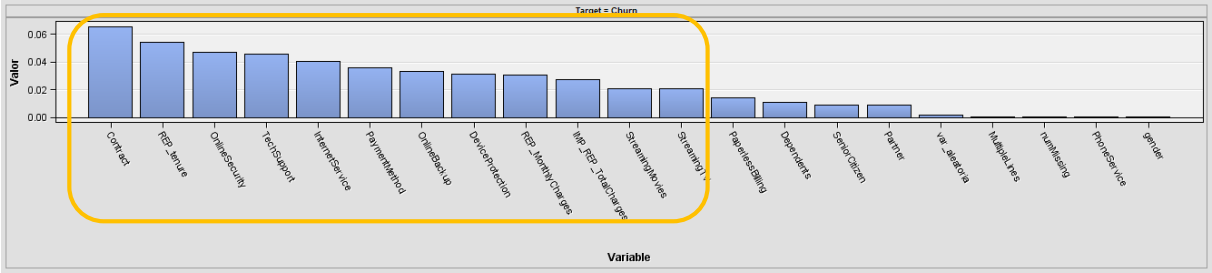


Figura 8 Grafico V de cramer

Transformación de variables

Se hace uso del nodo transformación de variables y en el método predeterminado para las variables input intervalo utilizamos “mejor” que hace uso del chi cuadrado y es el adecuado para las variables objetivo categórica.

Debido a que se tienen variables nominales las convertiremos en Dummys, posterior a esto conectamos el nodo metadatos para cambiar el rol de las nuevas variables que tendrán el nombre TI, con esto podremos ver la relación de las variables nominales con la variable objetivo.

Estadísticos de transformaciones

Fuente	Método	Nombre de la variable	Fórmula	Número de niveles	No ausente	Ausente	Mínimo	Máximo	Media	Desviación estándar	Asimetría	Curtosis	Etiqueta
Input	Original	Contract		3	.	0	.	.	.	.	.	.	Contract
Input	Original	Dependents		2	.	0	.	.	.	.	.	.	Dependents
Input	Original	DeviceProt...		3	.	0	.	.	.	.	.	.	DeviceProt...
Input	Original	IMP_REP_...		.	7043	0	19	8685	2282.75	2266.163	0.961933	-0.23072	IMP_REP_...
Input	Original	InternetServ...		3	.	0	.	.	.	.	.	.	InternetServ...
Input	Original	MultipleLines		3	.	0	.	.	.	.	.	.	MultipleLines
Input	Original	OnlineBack...		3	.	0	.	.	.	.	.	.	OnlineBack...
Input	Original	OnlineSecu...		3	.	0	.	.	.	.	.	.	OnlineSecu...
Input	Original	PaperlessB...		2	.	0	.	.	.	.	.	.	PaperlessB...
Input	Original	Partner		2	.	0	.	.	.	.	.	.	Partner
Input	Original	PaymentMe...		4	.	0	.	.	.	.	.	.	PaymentMe...
Input	Original	PhoneServi...		2	.	0	.	.	.	.	.	.	PhoneServi...
Input	Original	REP_Month...		.	7043	0	18	119	64.78305	30.0943	-0.22033	-1.25688	REP_Month...
Input	Original	REP_tenure		.	7043	0	0	72	32.37115	24.55948	0.23954	-1.38737	REP_tenure
Input	Original	SeniorCitizen		2	.	0	.	.	.	.	.	.	SeniorCitizen
Input	Original	Streaming...		3	.	0	.	.	.	.	.	.	Streaming...
Input	Original	StreamingTV		3	.	0	.	.	.	.	.	.	StreamingTV
Input	Original	TechSupport		3	.	0	.	.	.	.	.	.	TechSupport
Input	Original	gender		2	.	0	.	.	.	.	.	.	gender
Input	Original	var_aleatoria		.	7043	0	.0000502	0.999986	0.499718	0.287752	0.0068	-1.18369	var_aleatoria
Output	Computed	EXP_var_al...	exp(var_ale...	.	7043	0	1.00005	2.718244	1.717391	0.490808	0.355557	-1.05333	Transforme...
Output	Computed	LOG_REP_...	log(REP_te...	.	7043	0	0	4.290459	3.036873	1.15551	-0.77837	-0.63357	Transforme...
Output	Computed	OPT_IMP_...	Optimal Bin...	4	.	0	.	.	.	.	.	.	Transforme...
Output	Computed	OPT_REP_...	Optimal Bin...	4	.	0	.	.	.	.	.	.	Transforme...
Output	Computed	TI_Contract1	Dummy	2	.	0	.	.	.	.	.	.	Contract:Mo...
Output	Computed	TI_Contract2	Dummy	2	.	0	.	.	.	.	.	.	Contract:On...
Output	Computed	TI_Contract3	Dummy	2	.	0	.	.	.	.	.	.	Contract:Tw...
Output	Computed	TI_Depend...	Dummy	2	.	0	.	.	.	.	.	.	Dependent...
Output	Computed	TI_Depend...	Dummy	2	.	0	.	.	.	.	.	.	Dependent...
Output	Computed	TI_DeviceP...	Dummy	2	.	0	.	.	.	.	.	.	DeviceProt...
Output	Computed	TI_DeviceP...	Dummy	2	.	0	.	.	.	.	.	.	DeviceProt...
Output	Computed	TI_DeviceP...	Dummy	2	.	0	.	.	.	.	.	.	DeviceProt...
Output	Computed	TI_Internet...	Dummy	2	.	0	.	.	.	.	.	.	InternetServ...
Output	Computed	TI_Internet...	Dummy	2	.	0	.	.	.	.	.	.	InternetServ...
Output	Computed	TI_Internet...	Dummy	2	.	0	.	.	.	.	.	.	InternetServ...
Output	Computed	TI_Multiple...	Dummy	2	.	0	.	.	.	.	.	.	MultipleLin...

Tabla 10 Generación de dummys

Posterior a la transformación de variables podemos ver que las variables que menor información aportan a la variable objetivo son TI_Paymentmethod4, TI_Onlinebackup3, TI_Deviceprotection, TI_Streamingtv3, TI_streamingmovies3, EXP_Var_Aleatoria, TI_Multipllines1, NumMissing, TI_Multipllines2, TI_Phoneservice1, TI_Phoneservice2, TI_Gender1, TI_Gender2.

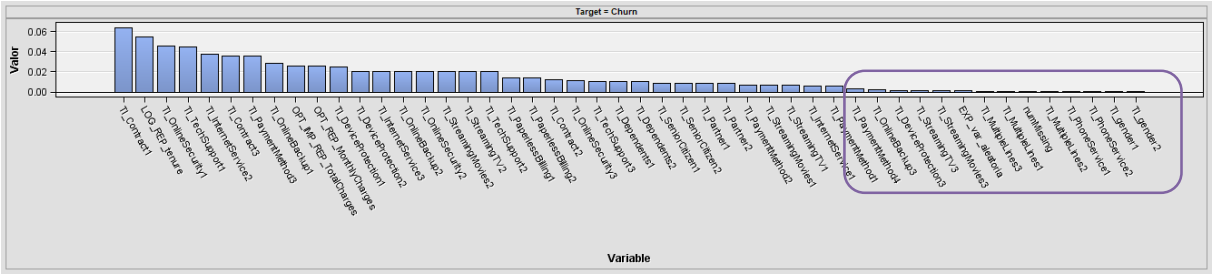


Figura 9 Figura Relación entre las variables

4. Modelización

4.1. Selección de variables

Finalizada la depuración de se extrae la base en formato SAS, en las diferentes herramientas (Miner, SAS) se ejecutarán los diferentes nodos y macros que nos permitan realizar una selección adecuada de las variables que serán tenidas en cuenta en el modelo.

Los resultados serán evaluados mediante un diagrama de cajas que nos permitirá ver su error y variabilidad.

Selección de variables E-MINER

Por medio de diferentes técnicas como regresión logística, selección de variables, mínimos cuadrados e incremento gradiente haremos uso de la base depurada y con el fin de saber cuál es la técnica que da mejores resultados se realiza un training test con 5 iteraciones con una partición de datos previa de 70/30 con método de partición estratificado.

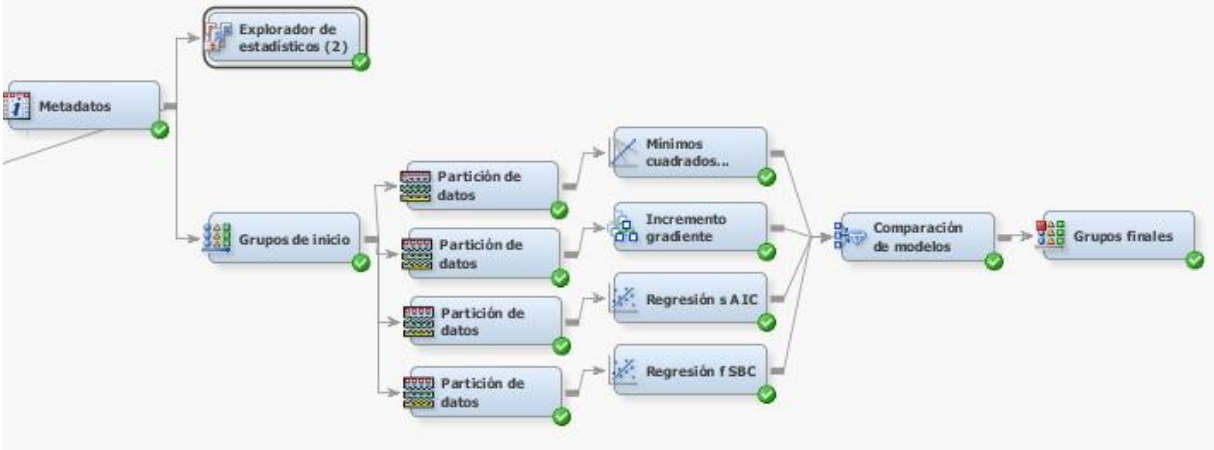


Diagrama 2 Caminos realizados para la selección de variables

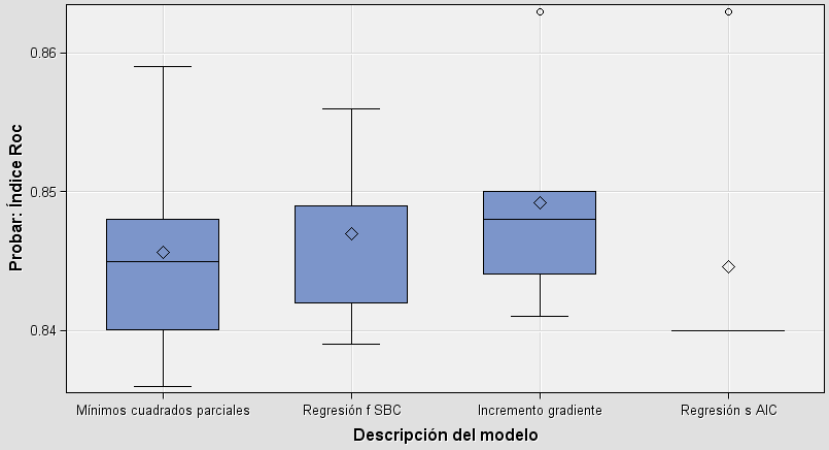


Figura 10 Selección de variables con Miner

En el gráfico de caja podemos observar que la variabilidad entre los modelos es similar, sin embargo, la dispersión de los datos en mayor en el modelo de mínimos cuadrados parciales.

Las variables seleccionadas por las diferentes técnicas son:

Nodo regresión PL (4)

La regresión de mínimos cuadrados nos ayudara a reducir los predictores a un conjunto de componentes menor no correlacionados y posterior a esto realiza una regresión de mínimos cuadrados sobre estos componentes, teniendo como resultado la selección de las siguientes variables: LOG_REP_tenure, OPT_IMP_REP_TotalCharges, TI_Contract1, TI_OnlineSecurity1, TI_TechSupport1.

Nodo Gradient Boosting (20)

Este algoritmo minimiza los residuos y permite obtener modelos con errores mínimos, controlando la constante de regularización y el número de iteraciones, las variables seleccionadas son: TI_Contract1, LOG_REP_tenure, TI_InternetService2, TI_OnlineSecurity1, TI_PaymentMethod3, TI_TechSupport1, OPT_REP_MonthlyCharges, TI_StreamingMovies3, OPT_IMP_REP_TotalCharges, TI_StreamingTV3, TI_MultipleLines1, TI_PaperlessBilling2, TI_OnlineBackup1, TI_PaperlessBilling1, TI_Contract2, TI_Contract3, TI_Dependents2, TI_SeniorCitizen2, EXP_var_aleatoria, TI_Partner1 .

Selección de variables con regresión stepwise –aic (14)

Este método de selección de variables se introducen todas las variables y se van excluyendo, eliminando la variable menos influyente, el criterio de selección trabajado es AIC, seleccionando las siguientes variables: LOG_REP_tenure, OPT_IMP_REP_TotalCharges, OPT_REP_MonthlyCharges, TI_Contract1, TI_Contract2, TI_InternetService2, TI_MultipleLines1, TI_OnlineSecurity1, TI_PaperlessBilling1, TI_PaymentMethod3, TI_SeniorCitizen1, TI_StreamingMovies3, TI_StreamingTV3, TI_TechSupport1.

Selección de variables con regresión forward –bsc (14)

En este método se introducen las variables de forma secuencial, iniciando con la variable con mayor correlación con la variable dependiente, las variables incluidas son aquellas que cumplen con el criterio de entrada, el criterio de selección fue bsc seleccionando las siguientes variables: LOG_REP_tenure, TI_Contract1, TI_Contract2, TI_InternetService2, TI_MultipleLines1, TI_OnlineSecurity1, TI_PaperlessBilling1, TI_PaymentMethod3, TI_StreamingMovies3, TI_StreamingTV3, TI_TechSupport1.

Estadísticos de ajuste

Agrupar índice	Grupo	Modelo seleccionado	Nodo predecesor	Nodo del modelo	Descripción del modelo	Variable target	Etiqueta target	Criterio de selección: Probar: Índice Roc	Entrenamiento: Error cuadrático medio
1		Y	PLS	PLS	Mínimos cuadrados pa...	Churn	Churn	0.859	0.136279
1			Reg3	Reg3	Regresión f SBC	Churn	Churn	0.842	0.131203
1			Boost	Boost	Incremento gradiente	Churn	Churn	0.841	0.131481
1			Reg	Reg	Regresión s AIC	Churn	Churn	0.84	0.128915
2		Y	Boost	Boost	Incremento gradiente	Churn	Churn	0.863	0.135597
2			Reg3	Reg3	Regresión f SBC	Churn	Churn	0.849	0.132052
2			Reg	Reg	Regresión s AIC	Churn	Churn	0.84	0.129272
2			PLS	PLS	Mínimos cuadrados pa...	Churn	Churn	0.836	0.13299
3		Y	Boost	Boost	Incremento gradiente	Churn	Churn	0.848	0.133437
3			PLS	PLS	Mínimos cuadrados pa...	Churn	Churn	0.848	0.134705
3			Reg	Reg	Regresión s AIC	Churn	Churn	0.84	0.128531
3			Reg3	Reg3	Regresión f SBC	Churn	Churn	0.839	0.13055
4		Y	Reg3	Reg3	Regresión f SBC	Churn	Churn	0.856	0.13395
4			Boost	Boost	Incremento gradiente	Churn	Churn	0.85	0.133671
4			PLS	PLS	Mínimos cuadrados pa...	Churn	Churn	0.845	0.134618
4			Reg	Reg	Regresión s AIC	Churn	Churn	0.84	0.129265
5		Y	Reg	Reg	Regresión s AIC	Churn	Churn	0.863	0.133129
5			Reg3	Reg3	Regresión f SBC	Churn	Churn	0.849	0.131892
5			Boost	Boost	Incremento gradiente	Churn	Churn	0.844	0.132498
5			PLS	PLS	Mínimos cuadrados pa...	Churn	Churn	0.84	0.133948

Tabla 11 Selección de variables regresión forward

Variables	Minimos Cuadrados	Gradient Boosting	Regresion Stepwise-AIC	Regresion Forward - BSC	Total
LOG_REP_tenure	x	x	x	x	4
TI_Contract1	x	x	x	x	4
TI_TechSupport1	x	x	x	x	4
OPT_IMP_REP_TotalCharges	x	x	x		3
TI_InternetService2		x	x	x	3
TI_OnlineSecurity1	x		x	x	3
TI_PaymentMethod3		x	x	x	3
TI_StreamingTV3		x	x	x	3
TI_MultipleLines1		x	x	x	3
TI_PaperlessBilling1		x	x	x	3
TI_Contract2		x	x	x	3
TI_StreamingMovies3		x	x	x	3
TI_OnlineBackup1		x			1
TI_PaperlessBilling2		x			1
TI_SeniorCitizen2		x			1
OPT_REP_MonthlyCharges		x	x		2
TI_Contract3		x			1
TI_Dependents2		x			1
EXP_var_aleatoria		x			1
TI_Partner1		x			1
TI_SeniorCitizen1			x		1
TI_OnlineSecurity1		x			1

Tabla 12 Resumen de selección de variables en Miner

La anterior selección se comparará con las variables seleccionadas con las macros en SAS, esto con el fin de consolidar las más usadas y generar tres sets de variables.

Selección de variables SAS

Se usa la macro randomselectlog con método stepwise repetidas veces la cual utiliza diferentes archivos train.

	efecto	Frequency Count	Percent of Total Frequency
1	TI_PaymentMethod3 TI_MultipleLines1 TI_OnlineSecurity1 TI_TechSupport1 TI_StreamingTV3 TI_SeniorCitizen1 TI_StreamingMovies3 TI_Contract2 TI_Contract1 LOG_REP_tenure OPT_IMP_REP_TotalCh OPT_REP_MonthI	21	37.5
2	TI_OnlineSecurity3 TI_PaymentMethod3 TI_MultipleLines1 TI_OnlineSecurity2 TI_TechSupport1 TI_StreamingTV3 TI_SeniorCitizen1 TI_StreamingMovies3 TI_Contract2 TI_Contract1 LOG_REP_tenure OPT_IMP_REP_Tot	7	12.5
3	TI_OnlineSecurity3 TI_PaymentMethod3 TI_MultipleLines1 TI_TechSupport1 TI_StreamingTV3 TI_SeniorCitizen1 TI_StreamingMovies3 TI_Contract2 TI_Contract1 LOG_REP_tenure OPT_IMP_REP_TotalCh OPT_REP_MonthI	5	8.9285714286

Tabla 13 Selección de variables en SAS - macro randomselectlog

Se hace uso del método de selección de varias proc logistic con método de selección stepwise.

	effect	modelo	DF	Wald Chi-square	Pr > Chi-Square	variable
1	TI_InternetService	TI_PaymentMethod3 TI_MultipleLines1 TI_OnlineSecurity1 TI_TechSupport1 TI_StreamingTV3 TI_SeniorCitizen1 TI_StreamingMovies3 TI_Contract2 TI_Contract1 LOG_REP_tenure OPT_IMP_REP_TotalCh OPT_REP_MonthlyChar TI_InternetService2	1	51.3022	<.0001	TI_PaymentMethod3 TI_MultipleLines1 TI_OnlineSecurity1 TI_TechSupport1 TI_StreamingTV3 TI_SeniorCitizen1 TI_StreamingMovies3 TI_Contract2 TI_Contract1 LOG_REP_tenure OPT_IMP_REP_TotalCh OPT_REP_MonthlyChar TI_InternetService2

Tabla 14 Selección de variables en SAS - Macro proc logistc

Posterior a la realización de los modelos de selección se ordenar en escala descendente de mayor a menor y se seleccionan 5 grupos para medir su efectividad:

Variables	Total	Variables	Total
LOG_REP_tenure	6	TI_MultipleLines1	5
TI_Contract1	6	TI_Contract2	5
TI_TechSupport1	6	TI_StreamingMovies3	5
OPT_IMP_REP_TotalCharges	5	TI_InternetService2	4
TI_OnlineSecurity1	5	OPT_REP_MonthlyCharges	4
TI_PaymentMethod3	5	TI_PaperlessBilling1	3
TI_StreamingTV3	5	TI_SeniorCitizen1	3

Tabla 15 Resumen selección de variables en Miner y SAS.

- Grupo 1: LOG_REP_tenure, TI_Contract1, TI_TechSupport1.
- Grupo 2: OPT_IMP_REP_TotalCharges, TI_OnlineSecurity1, TI_PaymentMethod3, TI_StreamingTV3, TI_MultipleLines1, TI_Contract2, TI_StreamingMovies3.
- Grupo 3: TI_InternetService2, TI_PaperlessBilling1, OPT_REP_MonthlyCharges, TI_SeniorCitizen1.
- Grupo 4 los mejores de cada grupo: LOG_REP_tenure, TI_Contract1, OPT_IMP_REP_TotalCharges, TI_OnlineSecurity1, TI_InternetService2, TI_PaperlessBilling1.
- Grupo 5 fusión de grupo 1 y 2: LOG_REP_tenure, TI_Contract1, TI_TechSupport1, OPT_IMP_REP_TotalCharges, TI_OnlineSecurity1, TI_PaymentMethod3, TI_StreamingTV3, TI_MultipleLines1, TI_Contract2, TI_StreamingMovies3.

4.2. Regresión logística

Regresión logística en SAS

Posterior a la selección de variables se elige el mejor modelo de regresión logística de acuerdo a los grupos anteriormente definidos, se hace uso de la macro cruzada logística, la cual hace uso de diferentes semillas lo cual nos permitirá elegir el mejor modelo de regresión.

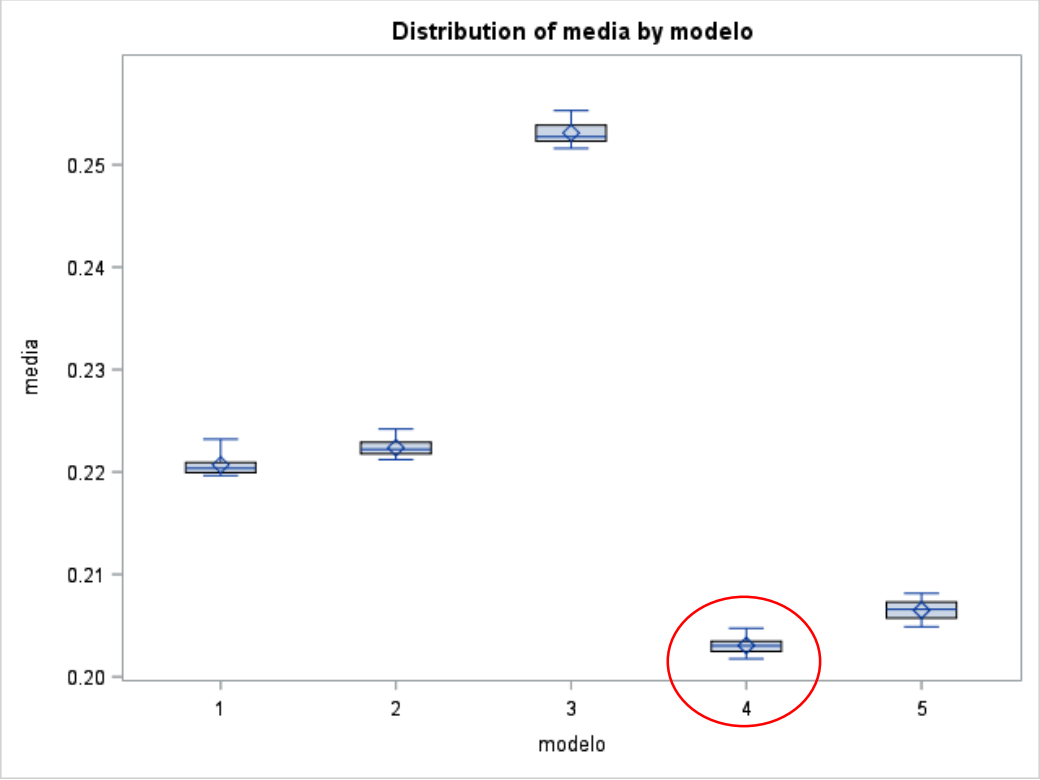


Figura 11 Diagrama de cajas regresión logística en SAS

El mejor modelo de regresión logística es el numero 4 el cual presenta poca variabilidad este corresponde al set de variables de los dos mejores de cada grupo, también el

modelo 5 podría ser un buen candidato este corresponde al set de variables fusión grupo 1 y 2.

Regresión logística en R

Se hace uso de la macro cruzada logística la cual usa las librerías dummis, mass, reshape, caret, dplyr y proc, la cual consiste en crear modelo de regresión logística usando especialmente la librería caret, creando las diferentes mediadas de evaluación del modelo como AUC y tasa de fallos y realizando validación cruzada de 5 repeticiones.

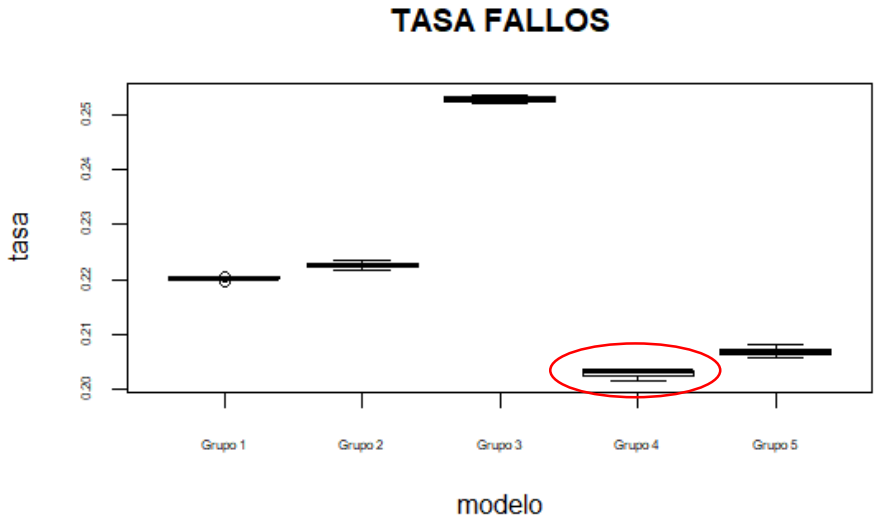


Figura 12 Diagrama de cajas regresión logística en R

Se observa en la figura 12 el modelo con el grupo de selección de variables 3 presenta la tasa de fallos más alta, al omitir este de la gráfica nos permite apreciar mejor los otros modelos, en el cual podemos ver en la figura 13 que el modelo Grupo 4 que corresponde al set de variables los dos mejores de cada grupo presenta poca mayor variabilidad respecto a los otros modelos pero la tasa de fallos es inferior , al igual que el sesgo es negativo, aunque otro buen candidato podría ser el grupo 5 que corresponde al set de variables unión de grupo 1 y 2.

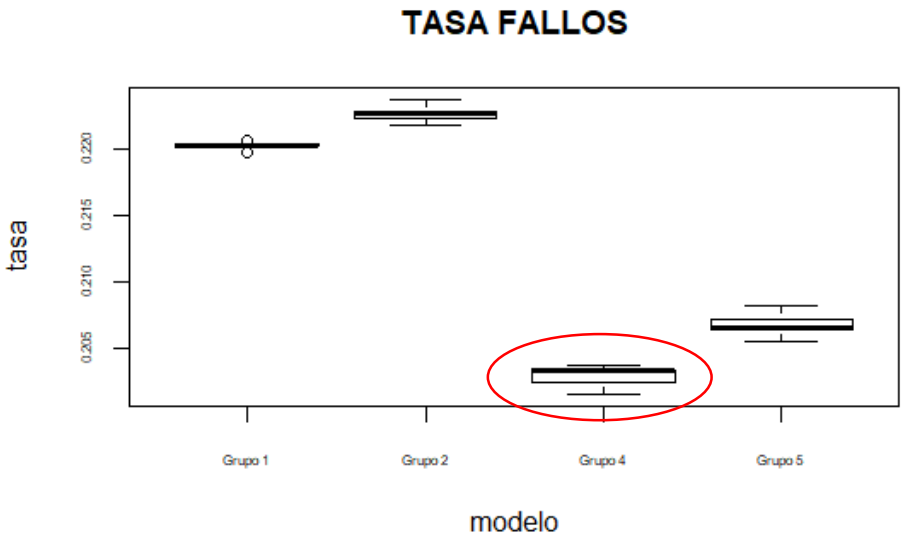


Figura 13 Diagrama de cajas regresión logística en R sin el grupo 3

Con respecto al uso de los dos softwares podemos ver que los resultados de los modelos de regresión lineal son muy similares, y en los dos casos dando como ganador al Grupo 4 que corresponde al set de variables los dos mejores de cada grupo.

4.3. Redes neuronales

Fijación del número de nodos

Teniendo en cuenta el número de variables input que se obtuvieron del método de selección de variables realizado en el paso anterior, se desea contar con un mínimo de 20 observaciones por parámetro y tener una idea inicial del número de nodos ocultos que permitan ajustar la curva, este fue calculado de acuerdo a la siguiente ecuación:

$$h + (k + 1) + h + 1$$

- Grupo 1: $h + (3 + 1) + h + 1 = 7043 \div 20 \quad h = 70$
- Grupo 2 : $h + (7 + 1) + h + 1 = 7043 \div 20 \quad h = 39$
- Grupo 3 : $h + (4 + 1) + h + 1 = 7043 \div 20 \quad h = 58$
- Dos mejores de cada grupo : $h + (6 + 1) + h + 1 = 7043 \div 20 \quad h = 44$
- Fusion de grupo 1 y 2 : $h + (10 + 1) + h + 1 = 7043 \div 20 \quad h = 29$

Construcción de las redes

Se generan las redes neuronales en SAS en las cuales el algoritmo optimiza los parámetros y obtiene las predicciones, para ellos se varían los siguientes parámetros:

- Variables: Con el fin de encontrar el mejor grupo de variables para el modelo se toman los grupos creados en el ítem anterior con el fin de realizar pruebas con ellos.

- Numero de nodos ocultos: El número de nodos se estable de acuerdo al cálculo realizado anteriormente sin embargo al realizar el tuneado de la red con la macro variar se identifica un sobre ajuste en la red por lo que se toma un numero de nodos inferior.

- Early stopping: Con el fin de evitar una red infra o sobre ajustada se usa early stopping con la macro redneuronalbinaria el cual divide los datos en training y validación, este detiene el proceso de estimación cuando error en los datos de validación empieza a aumentar.

- Método de activación: Se utiliza el método de activación softmax el cual es el adecuado para las redes neuronales de clasificación binaria.

- Algoritmo de optimización: Se varían los algoritmos de optimización Levmar y Bprop con cada grupo de variables, se verifica su comportamiento y se toman los mejores resultados de cada set de variables.

En la tabla 16 se relacionan las diferentes pruebas realizadas y sus resultados, esta tabla cuanta con los resultados más relevantes ya que por motivos de espacio no se publica completamente:

Grupo	N. de variables	Semilla	Nodos ocultos	Metodo de activacion	Metodo	Early Stopping	_VAVERR_
Grupo 5	10	442711	10	TANH	levmark	14	0.81863
Grupo 5	10	442711	7	TANH	levmark	11	0.82617
Grupo 4	6	442711	7	TANH	levmark	14	0.82865
Grupo 4	6	442713	10	TANH	levmark	10	0.83007
Grupo 4	6	442711	10	TANH	levmark	12	0.83138
Grupo 5	10	442713	7	TANH	levmark	15	0.83222
Grupo 4	6	442713	7	TANH	levmark	15	0.83586
Grupo 5	10	442713	10	TANH	levmark	10	0.83811
Grupo 5	10	442712	7	TANH	levmark	9	0.84475
Grupo 5	10	442711	7	TANH	Bprop	42	0.84623
Grupo 5	10	442711	10	TANH	Bprop	39	0.84669
Grupo 5	10	442713	10	TANH	Bprop	40	0.85597
Grupo 4	6	442713	10	TANH	Bprop	31	0.85627
Grupo 4	6	442713	7	TANH	Bprop	33	0.85629
Grupo 5	10	442713	7	TANH	Bprop	42	0.85634
Grupo 4	6	442711	7	TANH	Bprop	31	0.85783
Grupo 5	10	442712	10	TANH	levmark	9	0.85955
Grupo 4	6	442712	7	TANH	levmark	10	0.86004
Grupo 4	6	442711	10	TANH	Bprop	30	0.86246
Grupo 5	10	442712	7	TANH	Bprop	38	0.86350
Grupo 4	6	442712	10	TANH	levmark	9	0.86561
Grupo 1	3	442712	19	TANH	levmark	9	0.88860
Grupo 1	3	442712	13	TANH	levmark	11	0.88876
Grupo 2	7	442712	13	TANH	levmark	11	0.88876
Grupo 2	7	442712	9	TANH	levmark	11	0.88918
Grupo 2	7	442713	9	TANH	levmark	9	0.89572
Grupo 1	3	442713	13	TANH	levmark	9	0.89699
Grupo 2	7	442713	13	TANH	levmark	9	0.89699
Grupo 1	3	442713	19	TANH	levmark	9	0.89722
Grupo 1	3	442712	13	TANH	Bprop	35	0.90666
Grupo 5	10	442712	10	TANH	Bprop	22	0.90687
Grupo 1	3	442712	19	TANH	Bprop	35	0.90734
Grupo 1	3	442711	13	TANH	levmark	9	0.91281
Grupo 2	7	442711	13	TANH	levmark	9	0.91281
Grupo 2	7	442711	9	TANH	levmark	9	0.91285

Tabla 16 Tuneado de la red en SAS

Posterior a la identificación del número de nodos, el método, el algoritmo de optimización y el punto de corte del early stopping, se seleccionan las redes con mejores resultados de cada set de variables y se usa la macro cruzada binarias neural early usando varias semillas para elegir el modelo ganador, a continuación, se relacionan los resultados:

Grupo	N. de variables	Semilla	Nodos ocultos	Metodo de activacion	Metodo	Early Stopping	_VAVERR_
Grupo 1	3	442712	19	TANH	levmark	9	0.88860
Grupo 2	7	442712	13	TANH	levmark	11	0.88876
Grupo 3	4	442713	7	TANH	levmark	29	0.98021
Grupo 4	6	442711	7	TANH	levmark	14	0.82865
Grupo 5	10	442711	10	TANH	levmark	14	0.81863

Tabla 17 Redes con mejores resultados de acuerdo al set de variables

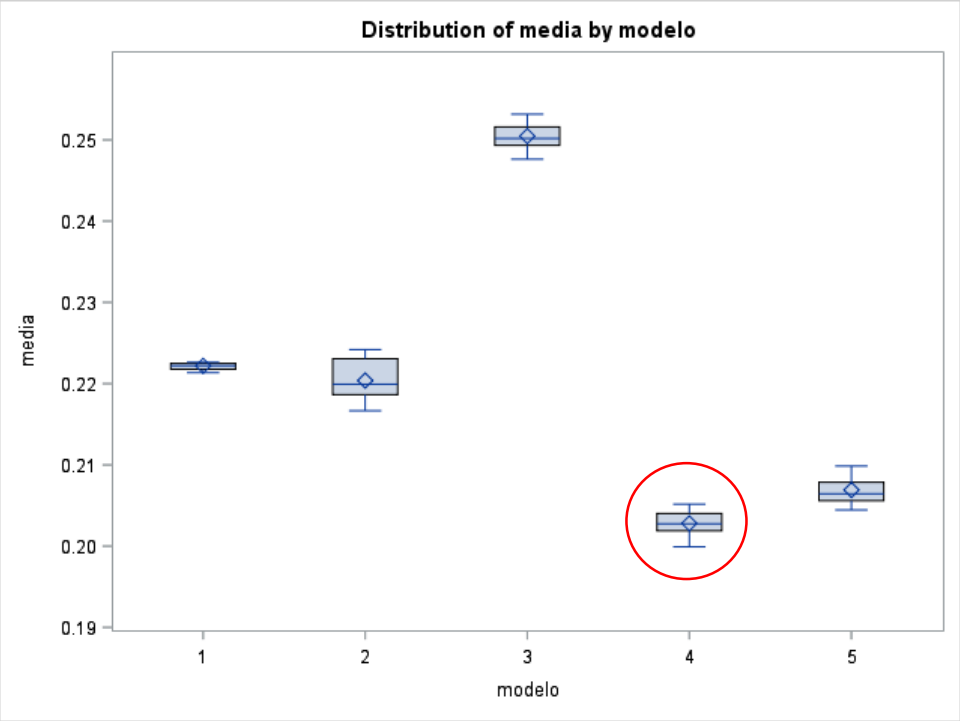


Figura 14 Diagrama de cajas de redes neuronales en SAS

De acuerdo al grafico de cajas la mejor red es la que pertenece al grupo 4 con el set de variables los dos mejores de cada grupo, vemos que no tiene mucha variabilidad, cuenta con 6 variables, 7 nodos, algoritmo de optimización levmark, y early stopping de 14.

Redes con R

Posterior a la identificación de los mejores sets de variables se realiza la prueba de estos sets con el software R, previamente a la generación de las redes se hace tuneado con la librería caret con el método avnnet, el cual nos permiten identificar el numero óptimo de nodos (size) y el learning rate (Deacay).

Se tendrá especial cuidado con numero bajos de learning rate ya que podrían encontrarse óptimos locales, para estos de iterara en diferentes ocasiones con el fin de garantizar que este error es realmente pequeño.

Tuning con avnnetgrid

```
size  decay  Accuracy   Kappa
   5  0.001  0.7903449  0.3728017
   5  0.010  0.7966491  0.4363675
   5  0.100  0.7965070  0.4417418
   7  0.001  0.7926738  0.3947485
   7  0.010  0.7966493  0.4428367
   7  0.100  0.7963647  0.4420882
  10  0.001  0.7880160  0.3759271
  10  0.010  0.7942354  0.4274759
  10  0.100  0.7951440  0.4382912
  13  0.001  0.7832457  0.3418291
  13  0.010  0.7915947  0.4236734
  13  0.100  0.7942354  0.4373072
  15  0.001  0.7887549  0.3791371
  15  0.010  0.7926168  0.4286476
  15  0.100  0.7933553  0.4305362
  17  0.001  0.7841817  0.3531843
  17  0.010  0.7906860  0.4281995
  17  0.100  0.7923613  0.4254278
  20  0.001  0.7824500  0.3472790
  20  0.010  0.7893798  0.4215604
  20  0.100  0.7917368  0.4319153

Tuning parameter 'bag' was held constant at a value of FALSE
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were size = 7, decay = 0.01 and bag = FALSE.
```

Usando el método avnnet se tienen los siguientes resultados:

Grupo	Size	Decay	Accuracy	Kappa
1	5	0.100	0.7792989	0.3627754
2	7	0.010	0.7828486	0.4018052
3	5	0.010	0.7506748	0.23809837
4	15	0.010	0.7984376	0.4455188
5	7	0.010	0.7966493	0.4428367

Tabla 18 Resultados de tuneado de la red con avnnetgrid en R

Posterior a la realización del tuneado y la identificación del número óptimo de nodos y learning rate se ejecuta la macro cruzadasavnnetbin por cada set de variables obteniendo los siguientes resultados:

Resultados de cruzada avnnetbin:

Grupo	Size	Decay	Bag	Accuracy	Kappa	AccuracySD	KappaSD
1	5	0.1	FALSE	0.777738	0.3586088	0.008814796	0.02704472
2	7	0.01	FALSE	0.7831056	0.4091817	0.008130256	0.02000564
3	5	0.01	FALSE	0.7509025	0.2434016	0.007210982	0.02222818
4	15	0.01	FALSE	0.7950887	0.4381663	0.008510478	0.02794187
5	7	0.01	FALSE	0.7951456	0.4412464	0.007835993	0.02068478

Tabla 19 Resultados de cruzada avnnetbin en R

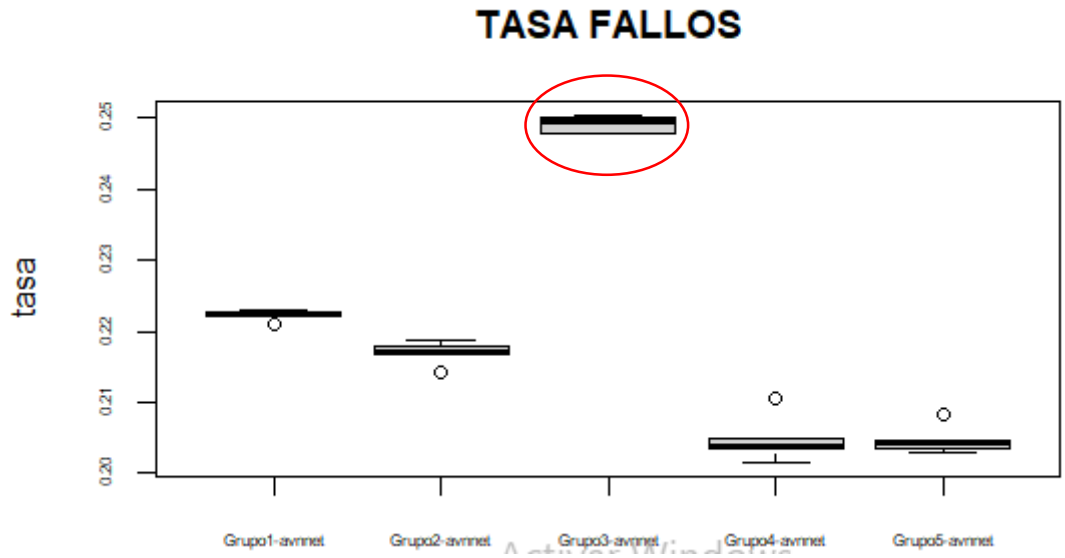


Figura 15 Diagrama de cajas en R de redes

Al observar la figura 15 se puede ver que la red del set de variables grupo 3 tiene una tasa de fallos superior a las demás, al igual que su variabilidad es mayor, se realiza de nuevo el grafico omitiendo está para escoger la red ganadora.

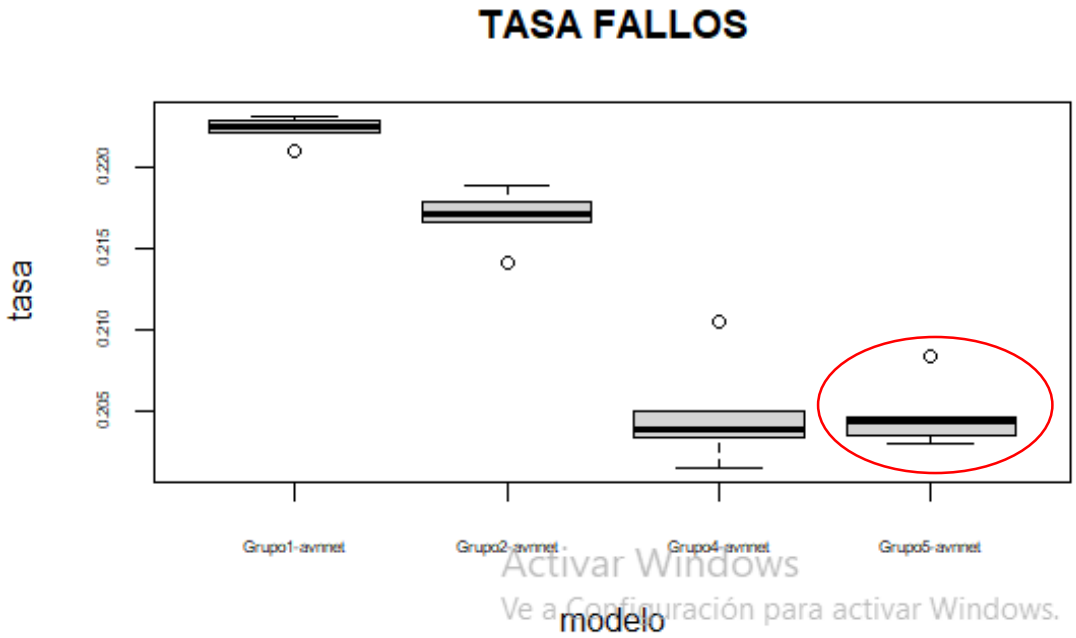


Figura 16 Diagrama de cajas tasa de fallos en R de redes – sin grupo 3

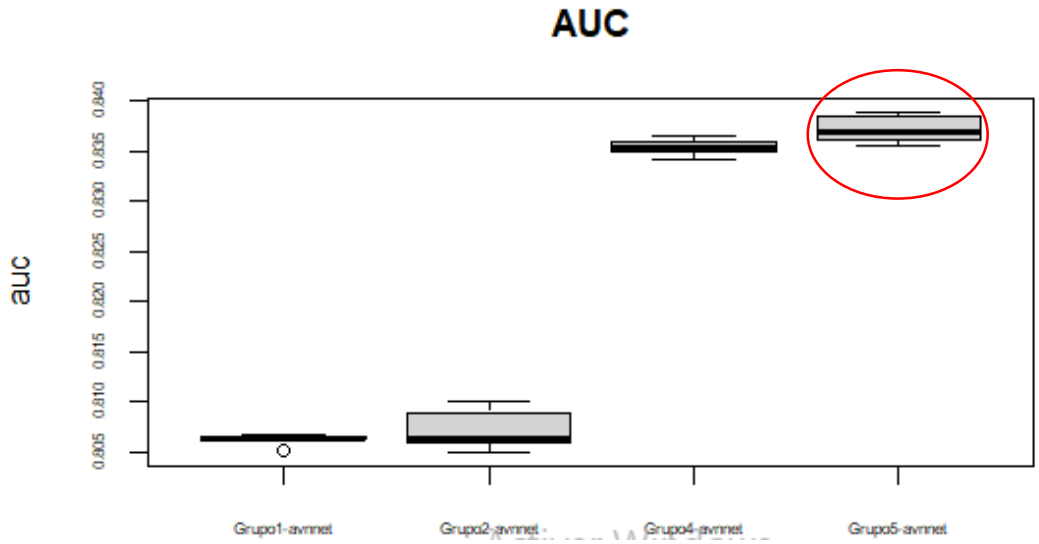


Figura 17 Diagrama de cajas AUC de redes en R

La red ganadora es la del grupo 5 del set de variables fusión grupo 1 y dos, tiene poca variabilidad y cuenta con 7 nodos y learnng rate de 0,01 un porcentaje de acierto (accurracy) de 0.795156, y un AUC superior a las otras redes.

Los resultados de los dos softwares son similares sin embargo en SAS se eligió como red ganadora el grupo 4 que pertenece al set de variables los dos mejores de cada grupo, y en R esta red con ese set de variables también tiene buenos resultados sin embargo fue superada por el grupo 5 que corresponde al set de variables fusión grupo 1 y 2 la cual tiene una menos variabilidad.

Arboles de clasificación

Los arboles de clasificación son algoritmos iterativos que permiten dividir los datos en regiones distintas, creando conjuntos homogéneos basados en la variable input más significativa, se crean valores constantes de acuerdo a cada región.

Con el fin de disminuir la función de error (tasa de fallos) se hace necesario identificar las regiones que permiten la reducción y evitar el sobre ajuste, para ello se puede variar el número de hojas finales, maxbranch, el p-valor y numero de ramas entre otros.

4.4. Bagging

Bagging con SAS

Haciendo uso de la macro cruzada ramdomforestbin se prueban los diferentes set de variables, se realizan pruebas variando el tamaño de hojas, profundidades de árbol y manteniendo el número de variables constante el cual corresponde al número total de variables por set de variables, durante el proceso de las pruebas se variaron el número

de iteraciones y el P-valor todo esto con el fin de controlar el sesgo y la varianza y obtener mejores modelos usando en este proceso validación cruzada y diferentes semillas.

Relaciono cuadro con los resultados más relevantes, donde los grupos corresponden a los diferentes sets de variables:

Grupo	Semilla	Maxtress	Variables	Porcenbag	Maxbranch	Tamhoja	Maxdepth	P-valor	Tasa de fallos
1	13345	100	3	0.80	4	25	5	0.1	0.22164
1	13345	100	3	0.80	4	15	4	0.1	0.22178
1	13347	100	3	0.80	4	25	5	0.1	0.22178
1	13345	100	3	0.80	4	15	4	0.5	0.22178
1	13349	100	3	0.80	4	15	4	0.1	0.22221
1	13349	100	3	0.80	4	25	5	0.1	0.22221
1	13349	100	3	0.80	4	15	4	0.5	0.22221
1	13347	100	3	0.80	4	15	4	0.5	0.22277
1	13348	100	3	0.80	4	15	4	0.1	0.22278
1	13347	100	3	0.80	4	15	4	0.1	0.22306
1	13348	100	3	0.80	4	15	4	0.5	0.22320
1	13346	100	3	0.80	4	15	4	0.1	0.22462
1	13346	100	3	0.80	4	15	4	0.5	0.22462
1	13348	100	3	0.80	4	25	5	0.1	0.22476
1	13345	100	3	0.80	4	5	10	0.1	0.2256124652
1	13346	100	3	0.80	4	25	5	0.1	0.22618
1	13347	100	3	0.80	4	5	10	0.1	0.2278853538
1	13348	100	3	0.80	4	5	10	0.1	0.2278853538
1	13349	100	3	0.80	4	5	10	0.1	0.2290235726
1	13346	100	3	0.80	4	5	10	0.1	0.2300149063
3	13347	100	4	0.80	4	15	4	0.5	0.30029
3	13348	100	4	0.80	4	15	4	0.1	0.30214
3	13347	100	4	0.80	4	15	4	0.1	0.30242
3	13347	100	4	0.80	4	25	5	0.1	0.30256
3	13346	100	4	0.80	4	15	4	0.5	0.30256
3	13346	100	4	0.80	4	15	4	0.1	0.30285
3	13346	100	4	0.80	4	25	5	0.1	0.30313
3	13348	100	4	0.80	4	15	4	0.5	0.30384
3	13349	100	4	0.80	4	15	4	0.5	0.30427
5	13346	200	10	0.80	4	8	3	0.5	0.30427
3	13349	100	4	0.80	4	15	4	0.1	0.30555
3	13348	100	4	0.80	4	25	5	0.1	0.30611
5	13349	200	10	0.80	4	8	3	0.5	0.30654
2	13347	100	7	0.80	4	15	4	0.1	0.30711
3	13345	100	4	0.80	4	15	4	0.1	0.30753
3	13345	100	4	0.80	4	25	5	0.1	0.30781
2	13347	100	7	0.80	4	25	5	0.1	0.30824
3	13345	100	4	0.80	4	15	4	0.5	0.30824
5	13349	100	10	0.80	4	15	4	0.1	0.30938
2	13345	100	7	0.80	4	25	5	0.1	0.30952
3	13349	100	4	0.80	4	25	5	0.1	0.31023

Tabla 20 Tuneado de parámetros de Bagging en SAS

Posterior a las pruebas realizadas en los árboles se identifican los mejores de cada set de variables los cuales relaciono a continuación:

Grupo	Modelo	Semilla	Maxtress	Variables	Porcenbag	Maxbranch	Tamhoja	Maxdepth	P-valor	Tasa de fallos
1	2	13345	100	3	0.80	4	15	4	0.1	0.22178
2	8	13345	100	7	0.80	4	15	4	0.5	0.31066
3	10	13345	100	4	0.80	4	15	4	0.1	0.30753
4	15	13345	100	6	0.80	4	25	5	0.1	0.31931
5	19	13346	100	10	0.80	4	25	5	0.1	0.32443

Grupo	Modelo	PorcenVN	PorcenFN	PorcenVP	PorcenFP	Sensi	Especif
1	2	0.6490630324	0.1555934128	0.1220897217	0.073253833	0.4396728016	0.8985849057
2	8	0.628052243	0.2067007382	0.0709823964	0.0942646224	0.2556237219	0.8694968553
3	10	0.6575809199	0.2561044861	0.0215786485	0.0647359455	0.0777096115	0.9103773585
4	15	0.6206700738	0.2055650199	0.0721181147	0.1016467916	0.2597137014	0.8592767296
5	19	0.597955707	0.1709256104	0.1067575241	0.1243611584	0.3844580777	0.8278301887

Tabla 21 Mejores resultados de tuneado de parámetros en Bagging por set de variables en SAS

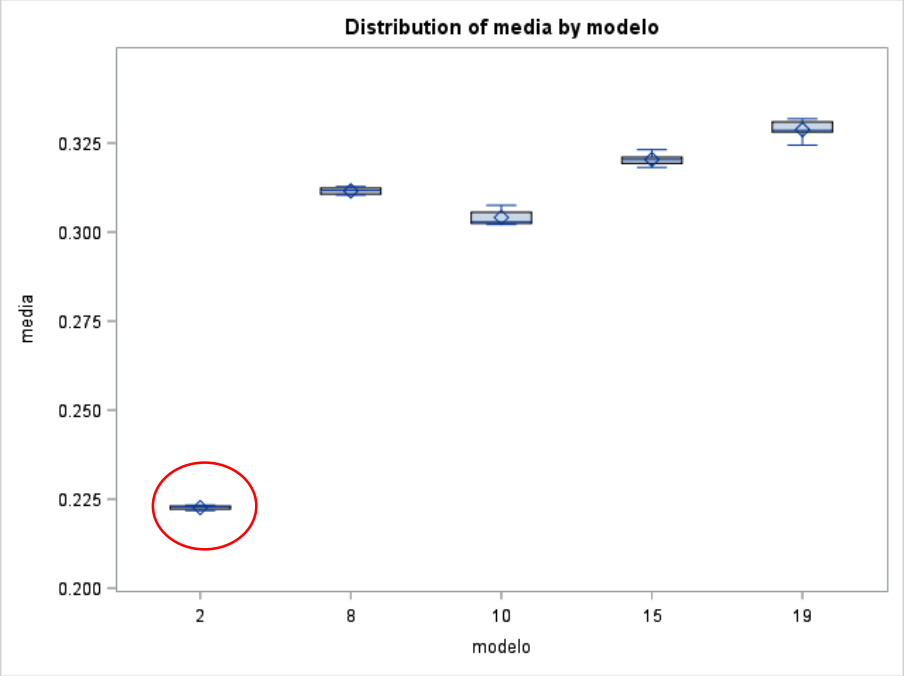


Figura 18 Diagrama de cajas Bagging en SAS

Observando el grafico de cajas anterior se identifica que el mejor modelo es el número 2 que corresponde al set de variables del grupo 1, teniendo poca varianza y una tasa de fallos baja, es un modelo sencillo que consta de tres variables, tamaño de hoja= 15, profundidad =4 y P- valor =0.1

Bagging con R

Se realiza el tuneado de los diferentes grupos de variables en los que se hace uso del archivo bagging en el cual se ingresa el parámetro MTRY de acuerdo al número total de variables de cada grupo, posterior se hace uso de la librería random forest y se plotea el error OOB a medida de avanzan las iteraciones este nos permite ver en que numero de árboles se estabiliza la tasa de fallos y de esta forma establecer un número de árboles a usar.

Se realiza el tuneado de la muestra a sortear (Sampsize), con los parámetros identificados en el tuneado se realiza la comprobación con validación cruzada con caret la cual nos permite ver si el Acurracy del árbol presenta una mejora, obteniendo los siguientes resultados:

Tuneado con caret:

Grupo	Accuracy	kappa	Mtry	Cp	Ntree	Sampsize
Grupo 1	0.7752375	0.3354466	3	0	3000	450
Grupo 2	0.7607543	0.3415337	7	0	2000	450
Grupo 3	0.7488297	0.2212969	4	0	3000	450
Grupo 4	0.7965366	0.419491	6	0	1000	450
Grupo 5	0.7963941	0.4164795	10	0	1000	450

Tabla 22 Resultado de Bagging con tuneado con caret en R

Resultados de validación cruzada con caret:

Grupo	Accuracy	kappa	Mtry
Grupo 1	0.7742428	0.3503118	3
Grupo 2	0.7829062	0.3708117	7
Grupo 3	0.7465547	0.2159776	4
Grupo 4	0.7942635	0.4193712	6
Grupo 5	0.7975292	0.4253332	10

Tabla 23 Resultado de Bagging con validación cruzada con caret en R

Para definir el árbol ganador se hace uso de la función cruzadarfbin la cual realiza validación cruzada de 5 repeticiones y se le establecen los parámetros tuneados anteriormente, arrojando los siguientes resultados:

Grupo	mtry	Accuracy	kappa	AcurracySD	KappaSD
Grupo 1	3	0.7668044	0.3470503	0.008542908	0.02351664
Grupo 2	7	0.7712626	0.3633028	0.00928476	0.02812956
Grupo 3	4	0.7380105	0.2272566	0.008965991	0.02166974
Grupo 4	6	0.7773964	0.3882483	0.007316648	0.02142996
Grupo 5	10	0.7844112	0.4012663	0.007978356	0.02258349

Tabla 24 Resultados Bagging con validación cruzada con función cruzadarfbin en R

Se puede observar que el porcentaje de acierto (Accurracy) es mayor en el grupo 5, seguido por el grupo 4, sin embargo, se muestra el diagrama de cajas que nos permitirá seleccionar el árbol ganador.

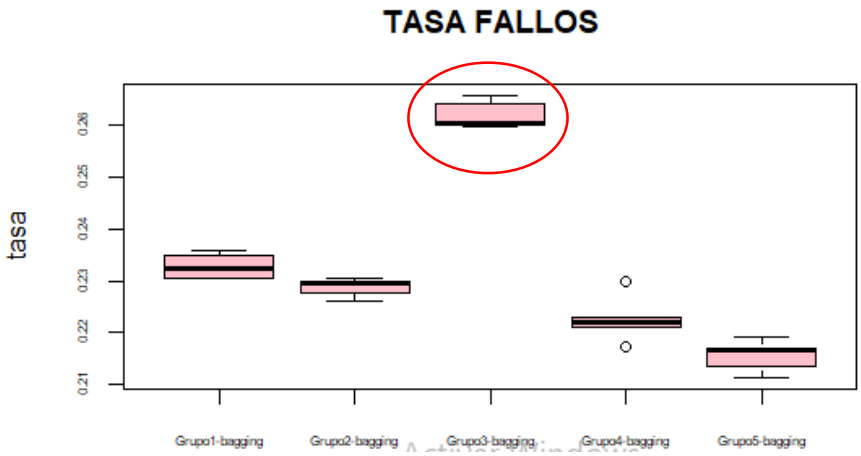


Figura 19 Diagrama de cajas- tasa de fallos Bagging en R

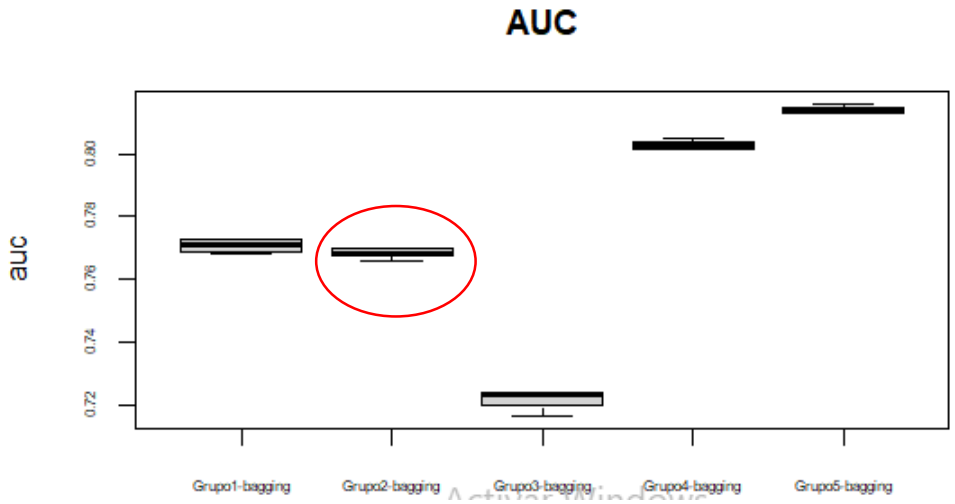


Figura 20 Diagrama de cajas- AUC Bagging en R

De acuerdo al diagrama de cajas se toma como modelo ganador el grupo 2 el cual presenta poca variabilidad, aunque el modelo 5 presenta una tasa de fallos más baja tiene mayor variabilidad, motivo por el cual no se toma como ganador, el modelo del grupo 2 cuenta con un porcentaje de aciertos del 0.7607543, mtry =7, número de árboles =2000.

4.5. Random forest

Random forest en SAS

Se hace uso de la macro randomforestbin para realizar el tuneado del árbol probando con diferentes valores de tamaño de hoja final, la profundidad máxima, el número de variables a sortear en cada nodo siempre menor al total de variables ya que no se desea realizar bagging, al realizar las pruebas también se valida con diferentes números de iteraciones y P-valor con el fin de evitar el sobre ajuste, el proceso se realiza con los diferentes set de variables y validación cruzada repetida y diferentes semillas.

Relaciono un resumen de los resultados de la validación cruzada:

Grupo	Semilla	Maxtress	Variables	Porcenbag	Maxbranch	Tamhoja	Maxdepth	P-valor	Tasa de fallos
1	13347	100	2	0.80	4	25	5	0.1	0.21965
1	13345	100	2	0.80	4	5	10	0.1	0.22164
1	13345	100	2	0.80	4	15	4	0.1	0.22164
1	13345	100	2	0.80	4	25	5	0.1	0.22164
1	13345	100	2	0.80	4	20	4	0.05	0.22164
1	13347	100	2	0.80	4	15	4	0.1	0.22178
1	13347	100	2	0.80	4	20	4	0.05	0.22178
1	13349	100	2	0.80	4	25	5	0.1	0.22221
1	13347	100	2	0.80	4	5	10	0.1	0.22277
1	13346	100	2	0.80	4	15	4	0.1	0.22277
1	13349	100	2	0.80	4	5	10	0.1	0.22292
1	13349	100	2	0.80	4	20	4	0.05	0.22292
1	13348	100	2	0.80	4	15	4	0.1	0.22320
1	13346	100	2	0.80	4	25	5	0.1	0.22320
1	13348	100	2	0.80	4	20	4	0.05	0.22320
1	13348	100	2	0.80	4	25	5	0.1	0.22363
1	13346	100	2	0.80	4	20	4	0.05	0.22419
1	13348	100	2	0.80	4	5	10	0.1	0.22420
1	13349	100	2	0.80	4	15	4	0.1	0.22462
1	13346	100	2	0.80	4	5	10	0.1	0.23087
3	13348	100	2	0.80	4	25	5	0.1	0.28113
3	13346	100	2	0.80	4	25	5	0.1	0.28312
3	13345	100	2	0.80	4	25	5	0.1	0.28567
3	13349	100	2	0.80	4	25	5	0.1	0.29092
3	13347	100	2	0.80	4	25	5	0.1	0.29106
3	13348	100	3	0.80	4	15	4	0.05	0.29192
3	13348	100	3	0.80	4	15	4	0.1	0.29220
3	13347	100	3	0.80	4	15	4	0.05	0.29546
3	13347	100	3	0.80	4	15	4	0.1	0.29617
3	13349	100	3	0.80	4	15	4	0.05	0.29930
3	13349	100	3	0.80	4	15	4	0.1	0.29987
3	13346	100	3	0.80	4	15	4	0.1	0.30171
3	13346	100	3	0.80	4	15	4	0.05	0.30200
3	13345	100	3	0.80	4	15	4	0.1	0.30327
3	13345	100	3	0.80	4	15	4	0.05	0.30512
2	13347	100	5	0.80	4	15	4	0.1	0.30782
5	13349	100	8	0.80	4	15	4	0.1	0.30796
2	13347	100	6	0.80	4	25	5	0.1	0.30810

Tabla 25 Resultado de tuneado random forest en SAS

Debido a que se tienen diferentes sets de variables se selecciona un ganador por cada set para luego compararlos y elegir el ganador, relaciono las características y resultados de los arboles seleccionados:

Grupo	Modelo	Semilla	Maxtress	Variables	Porcenbag	Maxbranch	Tamhoja	Maxdepth	P-valor	Tasa de fallos
1	24	13348	100	2	0.80	4	20	4	0.05	0.22320
2	25	13347	100	6	0.80	4	5	10	0.1	0.31946
3	30	13348	100	3	0.80	4	15	4	0.1	0.29220
4	34	13345	100	5	0.80	4	15	4	0.1	0.31349
5	40	13349	100	8	0.80	4	20	4	0.05	0.30810

Grupo	Modelo	PorcenVN	PorcenFN	PorcenVP	PorcenFP	Sensi	Especif
1	24	0.6490630324	0.1555934128	0.1220897217	0.073253833	0.4396728016	0.8985849057
2	25	0.6127200454	0.1964792731	0.0812038614	0.10959682	0.2924335378	0.8482704403
3	30	0.6581487791	0.2561044861	0.0215786485	0.0641680863	0.0777096115	0.911163522
4	34	0.6121521863	0.1998864282	0.0777967064	0.1101646792	0.2801635992	0.8474842767
5	40	0.6257808064	0.1885292447	0.0891538898	0.0965360591	0.3210633947	0.8663522013

Tabla 26 Mejores resultados de tuneado de parámetros en random forest por set de variables en SAS

Se realiza el grafico de cajas para observar su variabilidad y sesgo en el cual podemos ver que el modelo 24 que pertenece al set de variables Grupo 1 tiene una tasa de fallos baja respecto a los otros modelos al igual en términos de variabilidad, sin embargo, no se toma como ganador ya que puede tender al sobre ajuste, adicionalmente es un modelo muy sencillo que fue sorteado con solo 2 variables, como se ve a continuación:

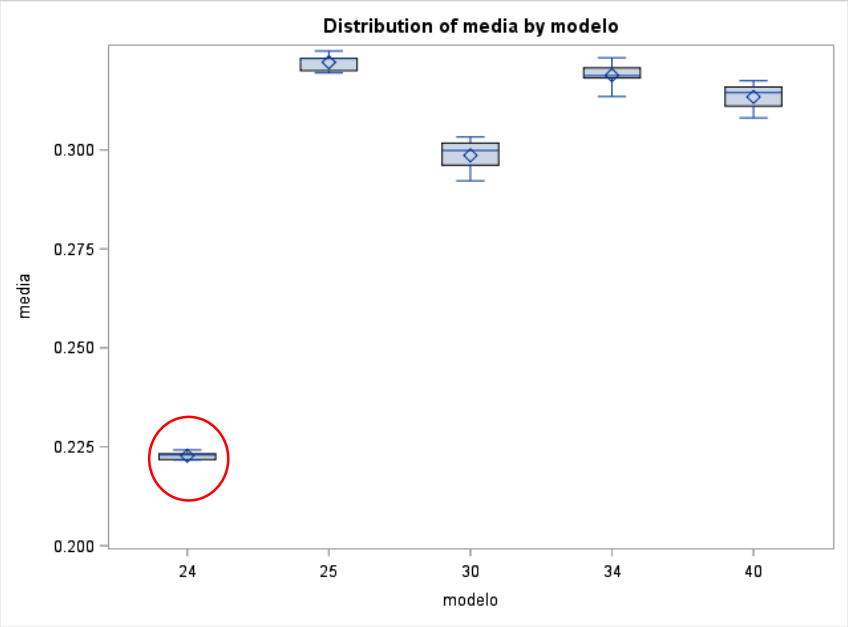


Figura 21 Diagrama de cajas random forest en SAS

Para la elección del modelo ganador se omite el modelo 24, para que nos permita ver con un poco más de detalle los otros modelos.

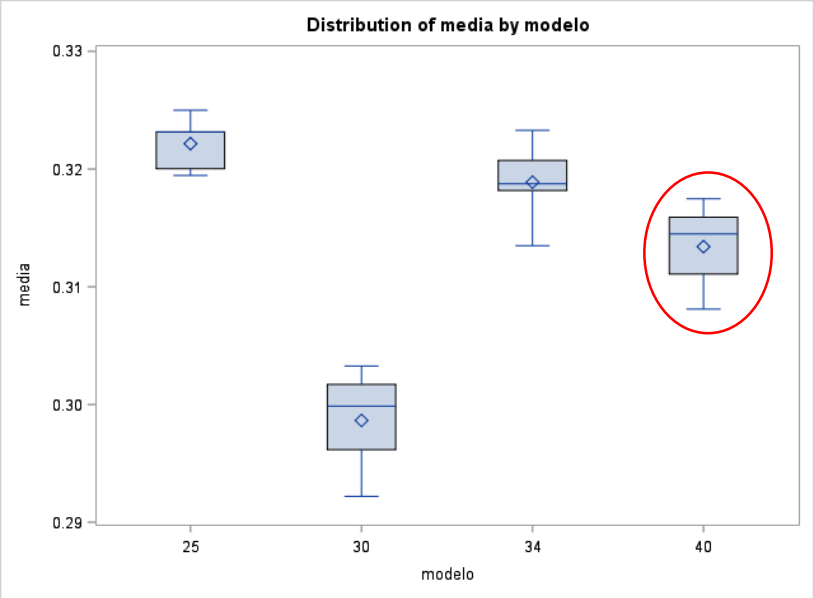


Figura 22 Diagrama de cajas random forest sin el modelo 24 en SAS

Se selecciona como ganador el modelo 40 que corresponde al set de variables del grupo 5,

Ya que presenta menor variabilidad y sesgo, este modelo cuenta con 8 variables, tamaño de hoja = 20, profundidad máxima = 4 y un P-valor de 0.05.

Random forest en R

Inicialmente se realiza el tuneado del parámetro MTRY con la librería caret, en el cual se prueban con varias opciones por cada set de variables, teniendo como máximo el número total de variables, seguido a esto se verifica la importancia de las variables en el modelo donde nos fijamos en el Mean Decrease Accurracy el cual nos indica en cuanto disminuye la predicción en caso de no incluir esa variable.

se hace uso de la librería random forest para plotear el error OOB a medida de avanzan las iteraciones este nos permite ver en que numero de árboles se estabiliza la tasa de fallos y de esta forma establecer un número de árboles para el modelo.

Se realiza el tuneado de la muestra a sortear (Sampsize), con los parámetros tuneados e identificados se realiza la comprobación validación cruzada con caret la cual nos permite ver si el Acurracy del árbol presenta una mejora, obteniendo los siguientes resultados:

Resultados de tuneados de parámetros Mtry, Sampsize y ploteado de OOB:

Grupo	Mtry	Accurracy	Kappa	Ntree	Sampsize
Grupo 1	2	0.7793536	0.3657481	600	450
Grupo 2	6	0.7807768	0.3637245	1000	450
Grupo 3	3	0.7485455	0.2076570	3000	450
Grupo 4	5	0.7982398	0.4235201	800	450
Grupo 5	9	0.7949749	0.4132041	2500	450

Tabla 27 Resultado de tuneado de random forest en R

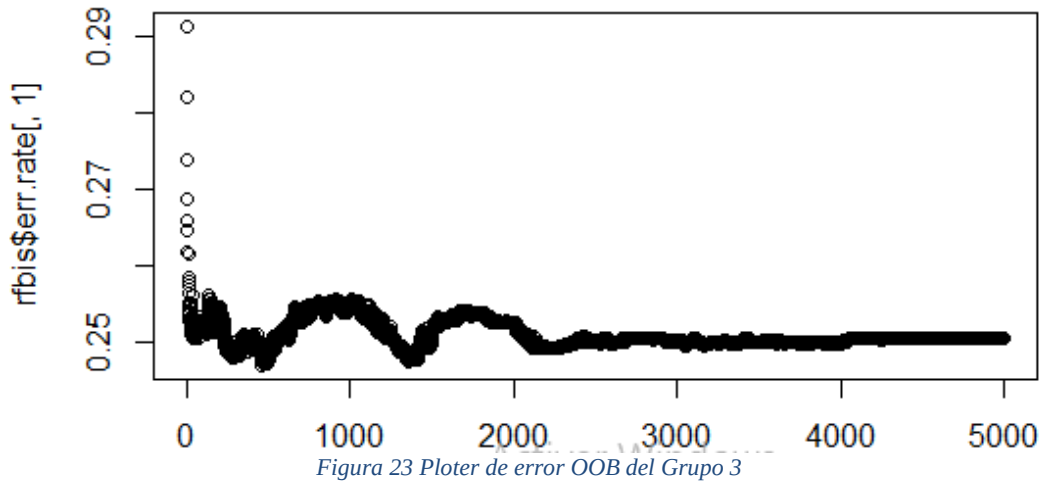


Figura 23 Ploter de error OOB del Grupo 3

Importancia de las variables: De acuerdo a los resultados del Grupo 2 no se incluye la variable TI_MultipleLines1, del Grupo 5 no se incluyen las variables TI_StreamingMovies3, TI_Contract2, para el resto de grupos se mantienen sus variables iniciales.

Grupo	variables	Mean Decrease Accuracy
Grupo_1	TI_TechSupport	45,94912
	LOG_REP_tenure	42,163510
	TI_Contract1	26,192310
Grupo_2	OPT_IMP_REP_TotalCharges	67,962940
	TI_OnlineSecurity1	53,533680
	TI_PaymentMethod3	30,083860
	TI_Contract2	21,229250
	TI_StreamingTV3	16,700010
	TI_StreamingMovies3	15,719450
	TI_MultipleLines1	11,694090
Grupo_3	TI_InternetService2	36,585660
	OPT_REP_MonthlyCharges	26,518810
	TI_SeniorCitizen1	17,990070
	TI_PaperlessBilling1	16,520000

Grupo	variables	Mean Decrease Accuracy
Grupo_4	LOG_REP_tenure	45,111760
	TI_InternetService2	30,240950
	TI_OnlineSecurity1	27,454630
	OPT_IMP_REP_TotalCharges	27,357870
	TI_Contract1	25,925270
	TI_PaperlessBilling1	17,043770
Grupo 5	LOG_REP_tenure	46,422290
	TI_Contract1	28,155590
	OPT_IMP_REP_TotalCharges	27,041010
	TI_OnlineSecurity1	24,098740
	TI_TechSupport1	23,535580
	TI_PaymentMethod3	18,911350
	TI_MultipleLines1	13,247570
	TI_StreamingTV3	12,745370
	TI_StreamingMovies3	10,520490
	TI_Contract2	-13,443000

Tabla 28 Importancia de las variables en random forest en R

Resultados de validación cruzada con caret:

Grupo	Accuracy	kappa	Mtry
Grupo 1	0.7755209	0.3516107	2
Grupo 2	0.7848943	0.377224	6
Grupo 3	0.7455654	0.210427	3
Grupo 4	0.7976721	0.4250044	5
Grupo 5	0.793981	0.4136869	9

Tabla 29 Resultados de validación cruzada random forest en R

Posterior a la realización del tuenado de los diferentes parámetros se hace uso de la función Cruzadarfbin que permiten plantear random forest, se realiza validación cruzada con 5 repeticiones y se elige el árbol ganador, a continuación, los resultados:

Grupo	Mtry	Accuracy	kappa	AcurracySD	KappaSD
Grupo 1	2	0.7762892	0.3537365	0.008455945	0.02585937
Grupo 2	6	0.7659522	0.350959	0.009740817	0.02879827
Grupo 3	3	0.7442858	0.2366212	0.006826251	0.02325237
Grupo 4	5	0.7798669	0.3952589	0.00729198	0.02005187
Grupo 5	9	0.7851494	0.4032696	0.007691176	0.02327251

Tabla 30 Resultados random forest de función cruzadabin en R

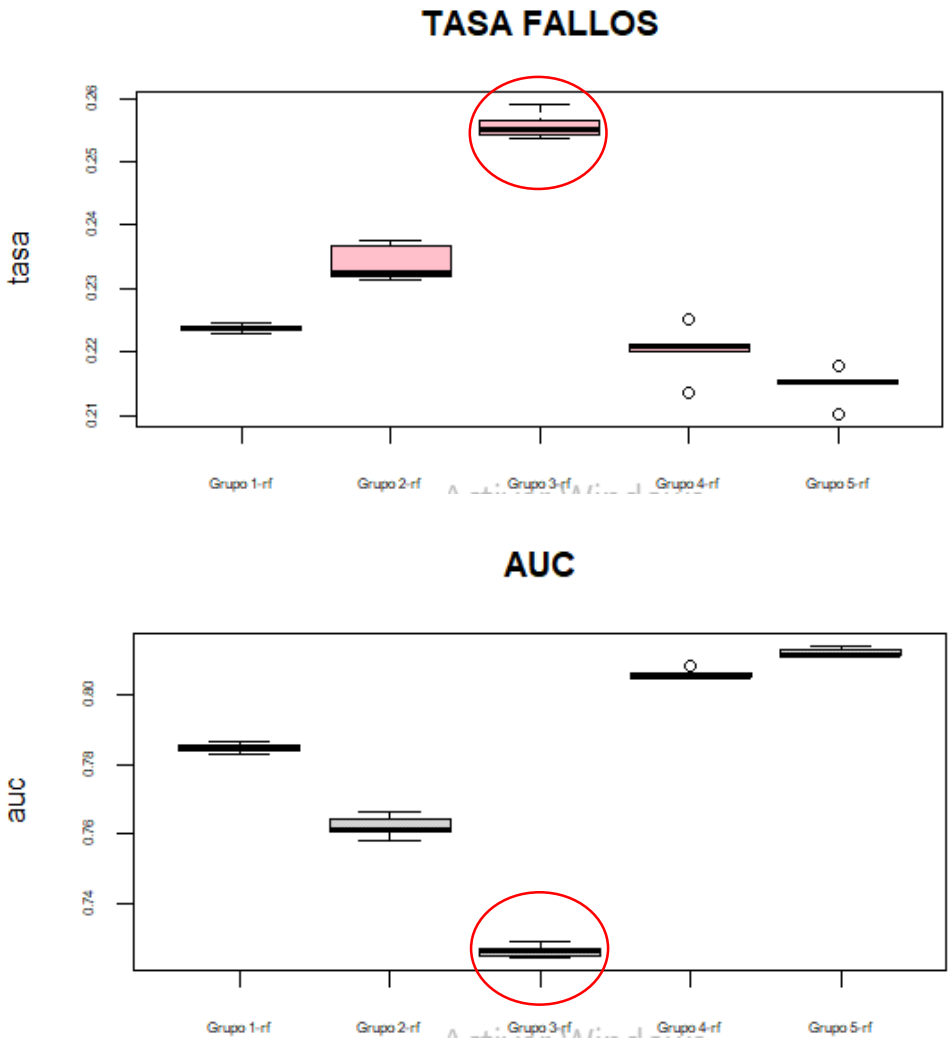


Figura 24 Diagrama de cajas random forest – tasa de fallos y AUC

En las gráficas anteriores podemos ver que el modelo del Grupo 3 presenta una tasa de fallos alta respecto a los demás al igual que su AUC es el más bajo, se realiza de nuevo el grafico omitiendo este grupo para seleccionar el ganador.

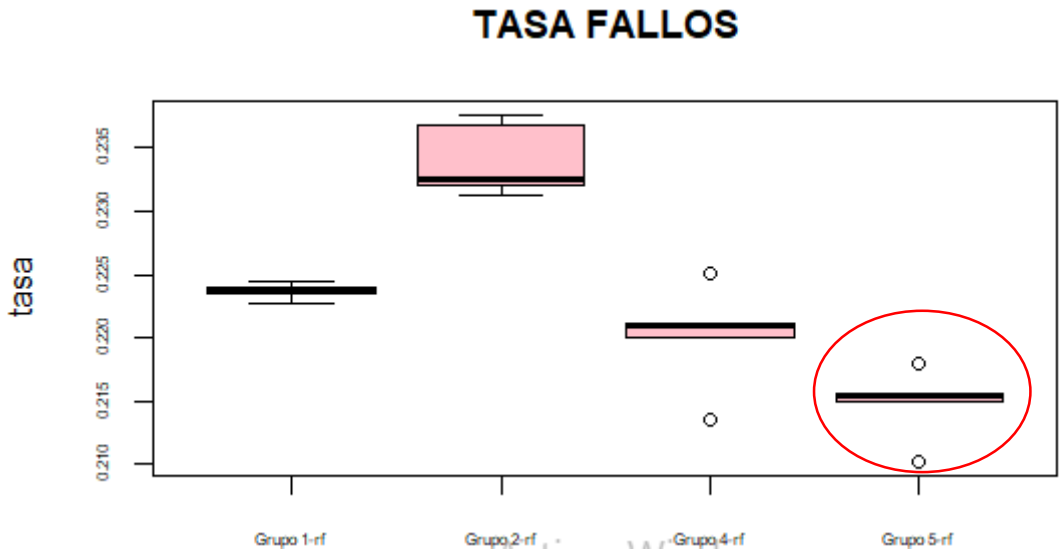


Figura 25 Diagrama de cajas random forest-omitiendo grupo 3

Se selecciona como ganador el Grupo 5 tiene una variabilidad pequeña y tasa de fallos menor, sin embargo, presenta algunos datos atípicos los cuales no superan el límite superior, este modelo pertenece al set de variables unión grupo 1 y 2, tiene un Mtry= 9, Ntree=2500 y un Accuracy = 0.7851494.

4.6. Gradient boosting

Gradient boosting con SAS

Este algoritmo de optimización actualiza la predicción de cada observación en la dirección del decrecimiento del error donde cada observación training se acerca al verdadero valor de Y, cuenta con las ventajas de los árboles, pero es más efectivo a la hora de reducir el error.

Al Igual que en los arboles anteriores se tunean diferentes parámetros como el parámetro de regularización, el número de iteraciones y el tamaño mínimo de la hoja final, se realiza el tuneado en los diferentes sets de variables, a continuación, relación un resumen de los resultados:

Grupo	Semilla	Leafsize	Iteraciones	Shrink	Maxbranch	Maxdepth	Mincatsize	Minobs	Tasa de fallos
5	13348	15	300	0.05	4	4	15	20	0.21383
5	13349	15	300	0.05	4	4	15	20	0.21482
5	13345	15	300	0.05	4	4	15	20	0.21823
5	13347	15	300	0.05	4	4	15	20	0.21922
4	13346	15	300	0.05	4	4	15	20	0.22064
5	13346	15	300	0.05	4	4	15	20	0.22064
5	13348	10	200	0.1	4	4	15	20	0.22292
4	13348	15	300	0.05	4	4	15	20	0.22334
5	13349	10	200	0.1	4	4	15	20	0.22363
5	13346	10	200	0.1	4	4	15	20	0.22405
4	13349	15	300	0.05	4	4	15	20	0.22490
2	13349	15	300	0.05	4	4	15	20	0.22491
4	13345	15	300	0.05	4	4	15	20	0.22504
4	13347	15	300	0.05	4	4	15	20	0.22547
4	13346	10	200	0.1	4	4	15	20	0.22618
2	13346	15	300	0.05	4	4	15	20	0.22632
5	13347	10	200	0.1	4	4	15	20	0.22647
2	13347	15	300	0.05	4	4	15	20	0.22789
1	13347	15	300	0.05	4	4	15	20	0.22831
1	13345	15	300	0.05	4	4	15	20	0.22845
1	13348	15	300	0.05	4	4	15	20	0.22860
5	13345	10	200	0.1	4	4	15	20	0.22874
1	13345	10	200	0.1	4	4	15	20	0.22916
1	13347	10	200	0.1	4	4	15	20	0.22945
2	13345	15	300	0.05	4	4	15	20	0.22945
2	13348	15	300	0.05	4	4	15	20	0.22973
1	13345	15	300	0.2	4	4	15	20	0.23016
4	13345	10	200	0.1	4	4	15	20	0.23129
1	13349	10	200	0.1	4	4	15	20	0.23144
1	13347	15	300	0.2	4	4	15	20	0.23158
4	13348	10	200	0.1	4	4	15	20	0.23158
1	13348	10	200	0.1	4	4	15	20	0.23172

Tabla 31 Resultado de tuneado de gradient boosting en SAS

Posterior al tuneado de los diferentes parámetros se seleccionan los modelos ganadores por cada set de variables para elegir el ganador, relaciono los resultados de los modelos seleccionados:

Grupo	Semilla	Leafsize	Iteraciones	Shrink	Maxbranch	Maxdepth	Mincatsize	Minobs	Tasa de fallos
1	13345	15	300	0.2	4	4	15	20	0.23016
2	13346	15	300	0.05	4	4	15	20	0.22632
3	13349	10	200	0.1	4	4	15	20	0.25784
4	13346	15	300	0.05	4	4	15	20	0.22064
5	13348	15	300	0.05	4	4	15	20	0.21383

Grupo	Modelo	PorcenVN	PorcenFN	PorcenVP	PorcenFP	Sensi	Especif
1	43	0.641113004	0.1544576945	0.1232254401	0.0812038614	0.4437627812	0.8875786164
2	44	0.644520159	0.1658148779	0.1118682567	0.0777967064	0.4028629857	0.8922955975
3	50	0.6519023282	0.2004542873	0.0772288472	0.0704145372	0.2781186094	0.9025157233
4	52	0.6269165247	0.1459398069	0.1317433277	0.0954003407	0.4744376278	0.8679245283
5	55	0.6342986939	0.136286201	0.1413969336	0.0880181715	0.509202454	0.8781446541

Tabla 32 Mejores resultados de tuneado por set de variables de gradient boosting

Se realiza un box plot para ver los modelos con más detalle en el cual se identifica que el modelo 50 que corresponde al grupo de variables 3 tiene una tasa de fallos superior a los demás, en cuanto al modelo 43 presenta poca variabilidad, pero se elige como ganador el modelo 52 que corresponde al grupo 4 set de variables que contiene los dos mejores de cada grupo, tiene una variabilidad y tasa de fallos inferior, es un modelo más complejo respecto al 43 en cuanto a su número de variables, los parámetros del modelo son tamaño mínimo de hoja final = 15, constante V= 0.05 y 300 iteraciones.

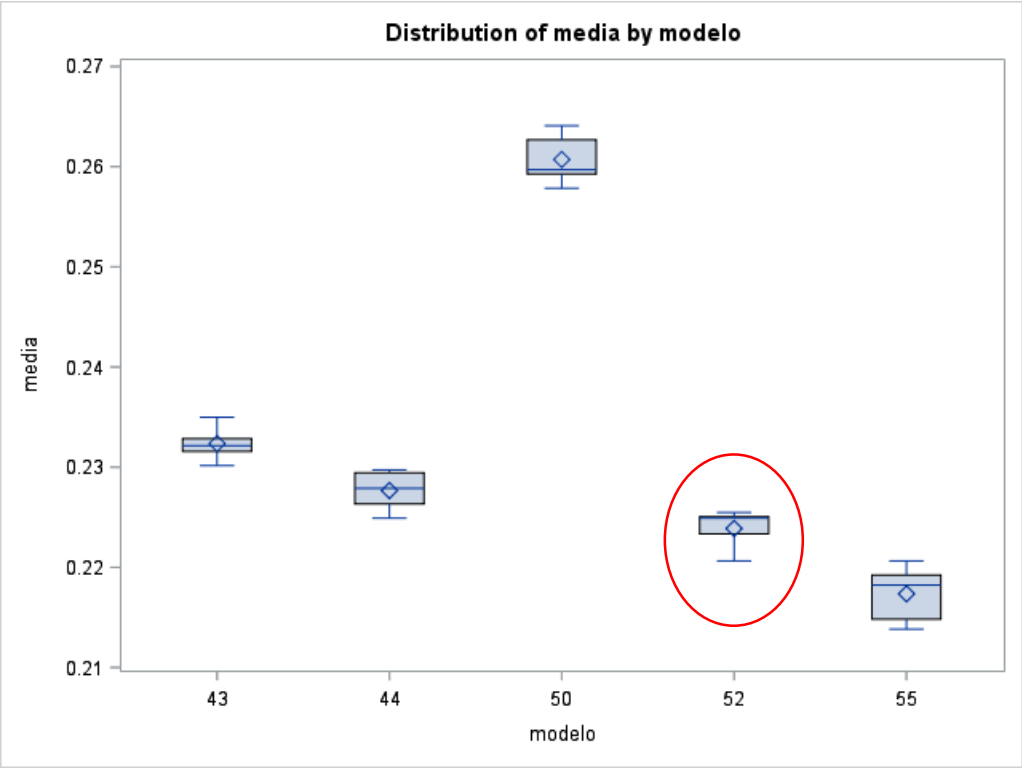


Figura 26 Diagrama de cajas gradient boosting en SAS

Gradient Boosting en R

Se realiza tuneado de cada grupo de variables con la librearía caret de los parámetros shrinekage que nos permitirá medir la velocidad del ajuste probando con valores entre 0.1 a 0.001, el tamaño mínimo de nodos finales (n.minobsinnode) nos permitirá ver la complejidad del árbol, si los valores son pequeños pueden ser arboles complejos y pueden sobre ajustar, se tunea con valores de 5 a 30.

Posterior al tuneado se realiza el estudio de Early stopping en el cual se fijan previamente los parámetros tuneados en la paso anterior y se verifica cómo evolucionan en función a las iteraciones (n.tree), fijándonos en el grafico o en los valores donde el accurracy alcanza su valor más alto respecto a las iteraciones, al igual que en los otros algoritmos de árboles se verifica la importancia de las variables y vemos el Mean Decrease Accurracy nos permite ver en cuanto disminuye la predicción en caso de no incluir la variable.

Resultados del tuneado y los parámetros seleccionados para caga grupo de variables:

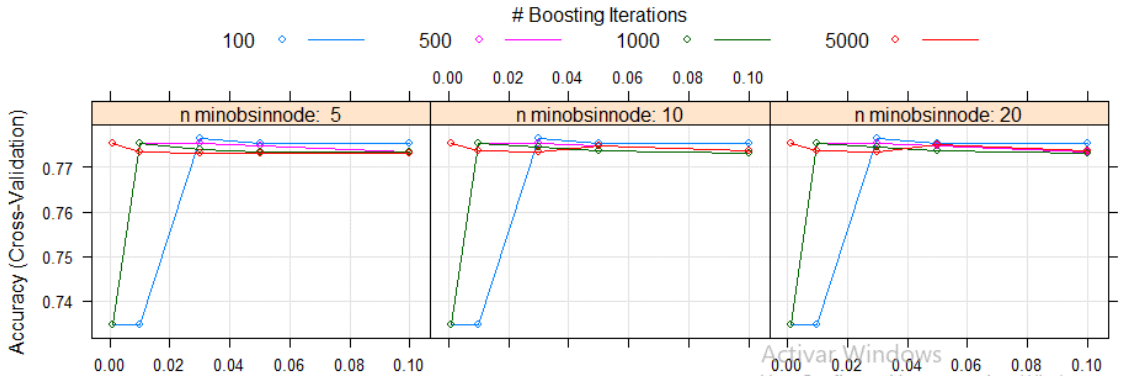


Figura 27 Gráfico de tuneado de parámetro shrinkage y el número de iteraciones del Grupo 1- gradient boosting en R

Grupo	N.tree	Interaction.depth	shrinkage	N.minobsinnode	Accurracy	Kappa
Grupo 1	100	2	0.03	5	0.7765154	0.3405997
Grupo 2	100	2	0.1	5	0.7817693	0.3757455
Grupo 3	1000	2	0.01	5	0.7503908	0.20383493
Grupo 4	500	2	0.05	20	0.7958260	0.4400699
Grupo 5	500	2	0.05	10	0.7966788	0.4374365

Tabla 33 Resultado de tuneado gradient boosting en R

Resultado de Early Stopping:

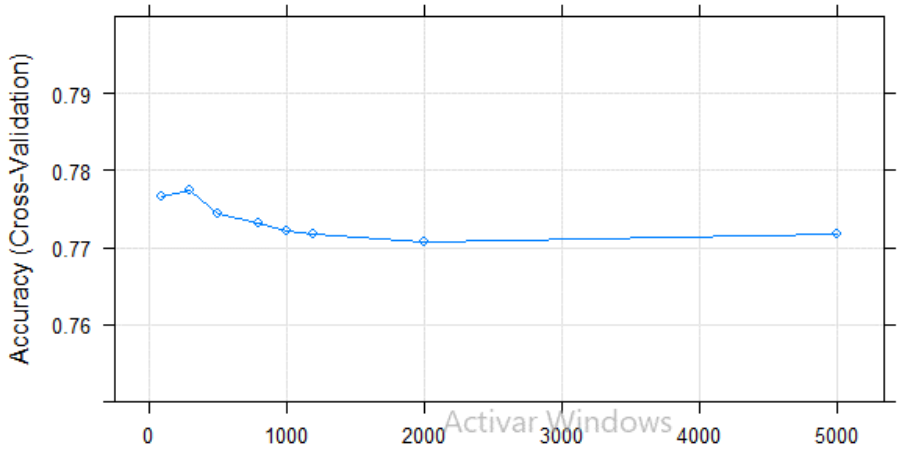


Figura 28 Resultado de Early Stopping del Grupo 1- gradient boosting en R

Grupo	N.tree	Accurracy	Kappa
Grupo 1	300	0.7773666	0.3695882
Grupo 2	100	0.7838994	0.3786171
Grupo 3	800	0.7498212	0.20086977
Grupo 4	300	0.7962525	0.4366562
Grupo 5	800	0.7962518	0.4378617

Tabla 34 Resultados de early stopping gradient boosting en R

Al ver los resultados de la importancia de las variables se eliminan las variables del Grupo 2: TI_StreamingTV3, TI_StreamingMovies3, TI_MultipleLines1, del Grupo 3: TI_SeniorCitizen1, del Grupo 4: OPT_IMP_REP_TotalCharges, TI_PaperlessBilling1, y para el Grupo 5: TI_MultipleLines1, TI_StreamingTV3, TI_Contract2 ya que al quitarlas la reducción en la predicción no se ve impactada.

Grupo	variables	Mean Decrease Accuracy
Grupo_1	TI_TechSupport	20,26419
	LOG_REP_tenure	19,70089
	TI_Contract1	60,03492
Grupo_2	OPT_IMP_REP_TotalCharges	33,652074
	TI_OnlineSecurity1	37,031004
	TI_PaymentMethod3	19,488615
	TI_Contract2	5,249503
	TI_StreamingTV3	1,648321
	TI_StreamingMovies3	1,842713
	TI_MultipleLines1	1,087770
Grupo_3	TI_InternetService2	56,803140
	OPT_REP_MonthlyCharges	31,918954
	TI_SeniorCitizen1	3,559765
	TI_PaperlessBilling1	7,718141

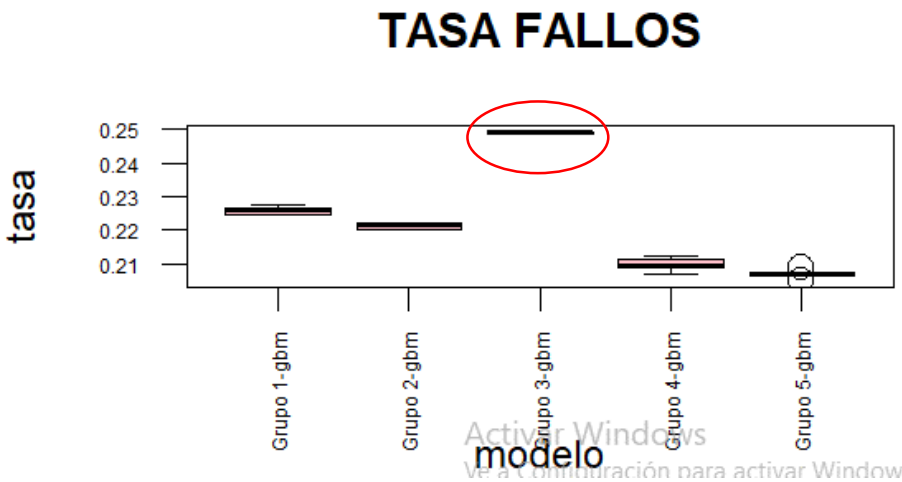
Grupo	variables	Mean Decrease Accuracy
Grupo_4	LOG_REP_tenure	22,049303
	TI_InternetService2	17,736985
	TI_OnlineSecurity1	12,132256
	OPT_IMP_REP_TotalCharges	5,035415
	TI_Contract1	40,638281
	TI_PaperlessBilling1	2,407759
Grupo 5	LOG_REP_tenure	19,23055
	TI_Contract1	39,40383
	OPT_IMP_REP_TotalCharges	8,15975
	TI_OnlineSecurity1	11,69799
	TI_TechSupport1	8,18601
	TI_PaymentMethod3	7,05569
	TI_MultipleLines1	1,25309
	TI_StreamingTV3	1,99937
	TI_StreamingMovies3	2,19517
	TI_Contract2	0,81855

Tabla 35 Importancia de las variables en gradient boosting en R

Una vez identificados los parámetros óptimos se usa la función Cruzadagbbin la cual realiza validación cruzada de 5 repeticiones, se incorporan los parámetros tuneados anteriormente y se generan los modelos para cada set de variables obteniendo los siguientes resultados:

Grupo	n.minobsinnode	shrinkage	n.trees	interaction.depth	Accuracy	Kappa	AccuracySD	KappaSD
1	5	0.03	300	2	0.7741312	0.3591695	0.008311416	0.02798279
2	5	0.1	100	2	0.778789	0.3667012	0.007791908	0.02404125
3	5	0.01	800	2	0.7509875	0.2199283	0.006985922	0.02235546
4	20	0.05	300	2	0.7901771	0.4085728	0.009093226	0.02832877
5	10	0.05	500	2	0.7930444	0.4300593	0.007136971	0.01906809

Tabla 36 Resultados función cruzadagbbin en gradient boosting en R



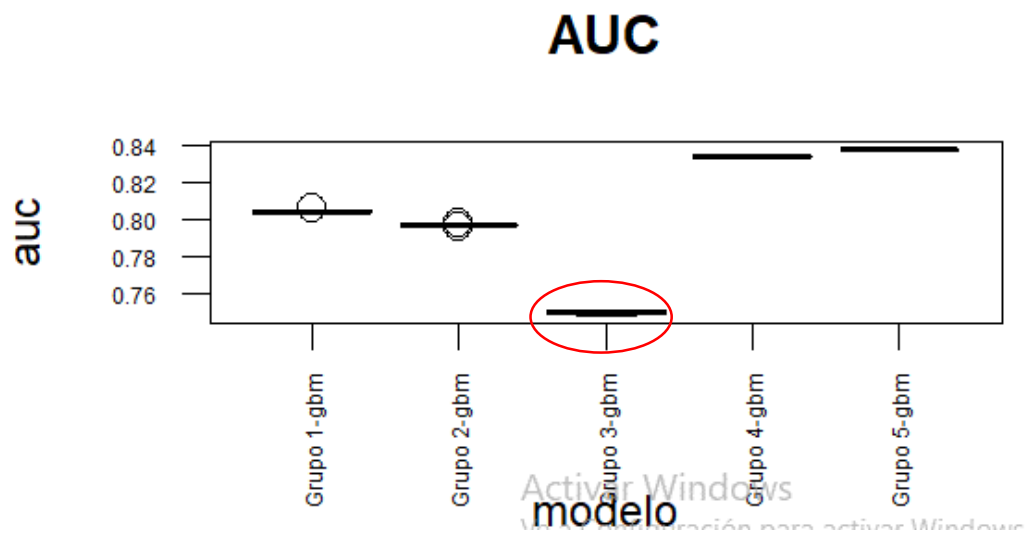


Figura 29 Diagrama de cajas gradient boosting en R- tasa de fallos y auc

De acuerdo a los resultados anteriores se puede evidenciar que los modelos tienen muy poca variabilidad y sesgo, el grupo 3 presenta la Tasa de fallos más alta al igual que el AUC más bajo respecto a los otros modelos, se omitirá este modelo y se graficará nuevamente para ver mejor los otros modelos y realizar la selección.

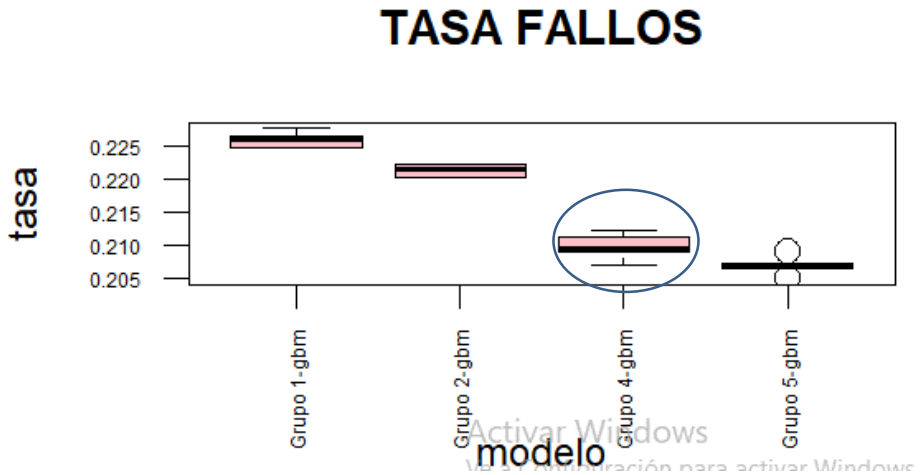


Figura 30 Diagrama de cajas gradient boosting- omitiendo grupo 3

Se selecciona el Grupo 4 como modelo ganador ya que cuanta con una tasa de fallos menor, poca variabilidad y sesgo, este modelo pertenece al set de variables unión grupo 1 y 2, tiene un n.minobsinnode= 20, shrinkage=0.05, n.tree= 500 y Acurracy =0.7930444.

4.7. SVM

SVM en SAS

Se realiza la prueba con los 5 set de variables y los diferentes kernels, el kernel polynomial y rbf en SAS presentaba errores en los cuales ejecutaba el proceso, pero no tenía resultados, sin embargo, con el parámetro kernel lineal funcionaba bien, se probó alternando el número la contante C de regularización del margen probando distintos

valores tratando de tener menores residuos y evitando el sobre ajuste y un grado de polinomio 1.

Sin embargo, por set de variables el resultado era el mismo solo presentaba diferencia cuando tomaba otros sets variables, como se ve en la tabla cruzadasvmbin y en la ilustración de svm.

	modelo	media	semilla
1	Grupo1-59-SVM-lineal	0.2266	12345
2	Grupo1-59-SVM-lineal	0.2258	12346
3	Grupo1-59-SVM-lineal	0.2258	12347
4	Grupo1-59-SVM-lineal	0.2268	12348
5	Grupo1-59-SVM-lineal	0.2258	12349
6	Grupo1-60-SVM-lineal	0.2266	12345
7	Grupo1-60-SVM-lineal	0.2258	12346
8	Grupo1-60-SVM-lineal	0.2258	12347
9	Grupo1-60-SVM-lineal	0.2268	12348
10	Grupo1-60-SVM-lineal	0.2258	12349
11	Grupo1-61-SVM-lineal	0.2266	12345
12	Grupo1-61-SVM-lineal	0.2258	12346
13	Grupo1-61-SVM-lineal	0.2258	12347
14	Grupo1-61-SVM-lineal	0.2268	12348
15	Grupo1-61-SVM-lineal	0.2258	12349
16	Grupo2-62-SVM-lineal	0.2358	12345
17	Grupo2-62-SVM-lineal	0.2258	12346
18	Grupo2-62-SVM-lineal	0.2626	12347
19	Grupo2-62-SVM-lineal	0.2466	12348
20	Grupo2-62-SVM-lineal	0.2626	12349
21	Grupo2-63-SVM-lineal	0.2358	12345
22	Grupo2-63-SVM-lineal	0.2258	12346
23	Grupo2-63-SVM-lineal	0.2626	12347
24	Grupo2-63-SVM-lineal	0.2466	12348
25	Grupo2-63-SVM-lineal	0.2626	12349
26	Grupo3-66-SVM-lineal	0.2626	12345
27	Grupo3-66-SVM-lineal	0.2626	12346
28	Grupo3-66-SVM-lineal	0.2626	12347
29	Grupo3-66-SVM-lineal	0.2626	12348

Tabla 37 Resultadoscruzadasvmbin SVM en SAS

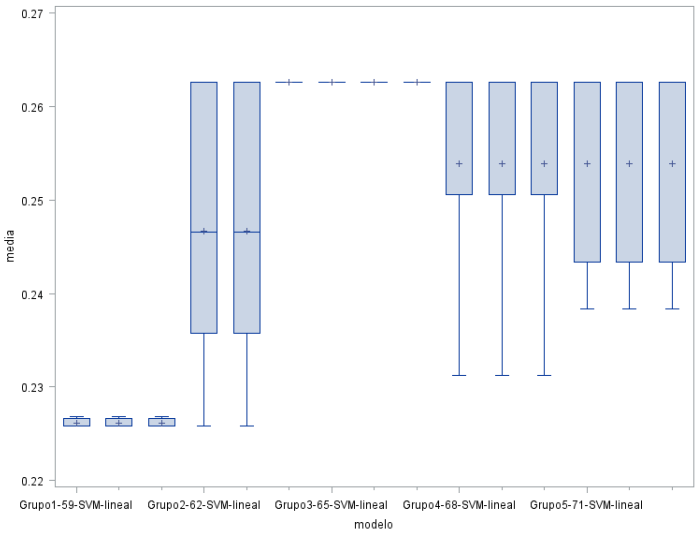


Figura 31 Diagrama de cajas SVM en SAS

Partiendo de que los resultados son los mismo por set de variables se seleccionan los modelos más sencillos, quedando los siguientes:

Modelo	Media
Grupo1-60-SVM-lineal	0.2258
Grupo2-63-SVM-lineal	0.2258
Grupo3-66-SVM-lineal	0.2626
Grupo4-69-SVM-lineal	0.2312
Grupo5-72-SVM-lineal	0.2384

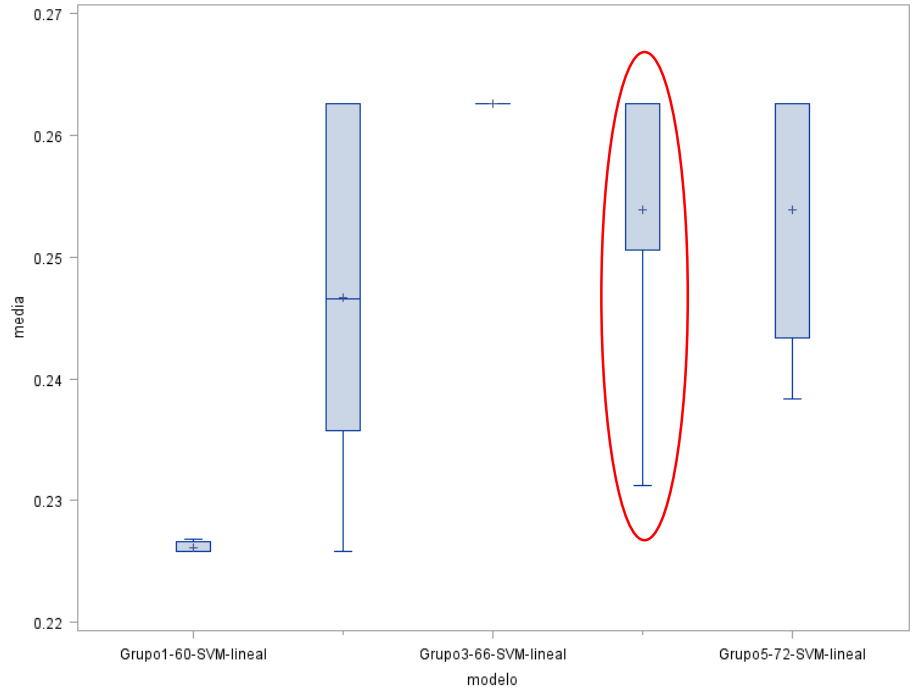


Figura 32 Diagrama de cajas SVM en SAS modelos seleccionados

De acuerdo a los resultados se toma como modelo ganador el Grupo 4-69_SVM-lineal que pertenece al set de variables los dos mejores de cada grupo, el cual tiene una variabilidad menor respecto a los otros modelos, pero una dispersión mayor, no se toma como ganador el modelo del Grupo 1 ya que es un modelo muy sencillo que tan solo consta de 3 variables.

SVM en R

Se realiza el tuneado con los diferentes set de variables y se prueban los diferentes funciones de kernel en cada set, se inicia con SVM lineal el cual permite maximizar el margen de separación, en este kernel solo se tunea la constante de regularización C, teniendo presente que a mayor C se obtiene un menor sesgo y más ajuste, se prueban distintos valores desde 0.01 hasta 10, posterior a este tuneado se observa el plot en el cual se intenta identificar algún patrón y se correo de nuevo la rejilla en esta región y se

selecciona el parámetro C que presenta mayor Accurracy teniendo los siguientes resultados:

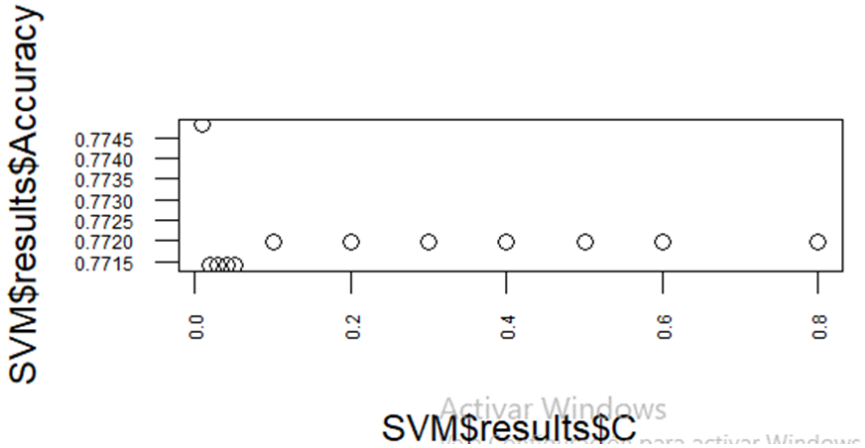


Figura 33 Plot de tuneado de parámetro C en Grupo 1-SVM en R

Grupo	C	Accurracy	Kappa	AccurracySD	KappaSD
Grupo 1	0.01	0.7748156	0.3539792	0.010384688	0.01859980
Grupo 2	0.02	0.7777943	0.3727229	0.007485502	0.02063836
Grupo 3	0.01	0.7346301	0	0.0002683866	0
Grupo 4	0.03	0.7915662	0.4068199	0.008582021	0.03018028
Grupo 5	0.2	0.7932689	0.4235891	0.008314575	0.02605178

Tabla 38 Resultado de tuneado en SVM lineal en R

Se realiza el tuneado con SVM Polinomial esta función busca extenderse a separaciones no lineales , al igual que en la función anterior se tunea el parámetro C, el grado del polinomio probando con valores de 2 y 3 y la Scale tomando valores entre 0.1 y 5, seguido a esto se genera un plot de las tres variables y se observan los primero patrones del grafico en especial del parámetro Degree, identificado este parámetro se realiza un plot de los otros dos parámetros y definimos los valores a tomar de Scale y C, como se ve a continuación los resultados por cada grupo de variables:

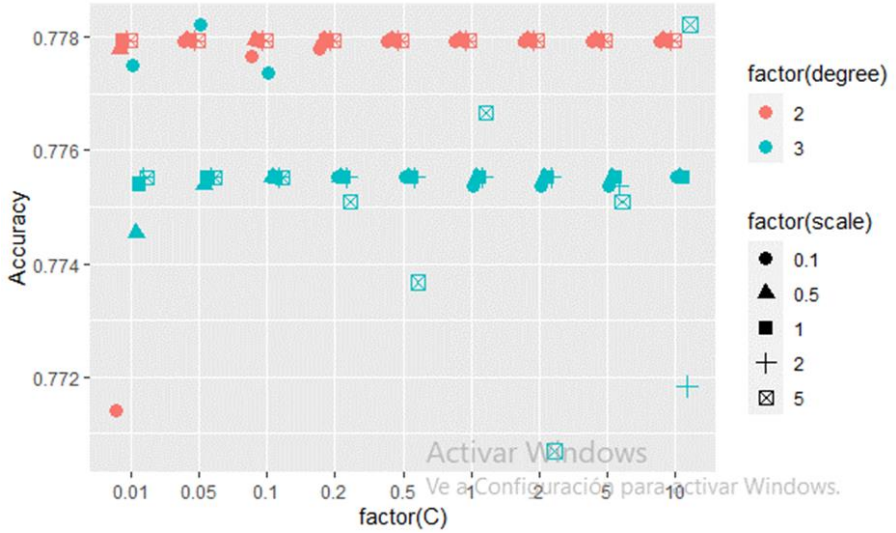


Figura 34 Plot 1 definición de parámetro Degree Grupo 1-SVM en R

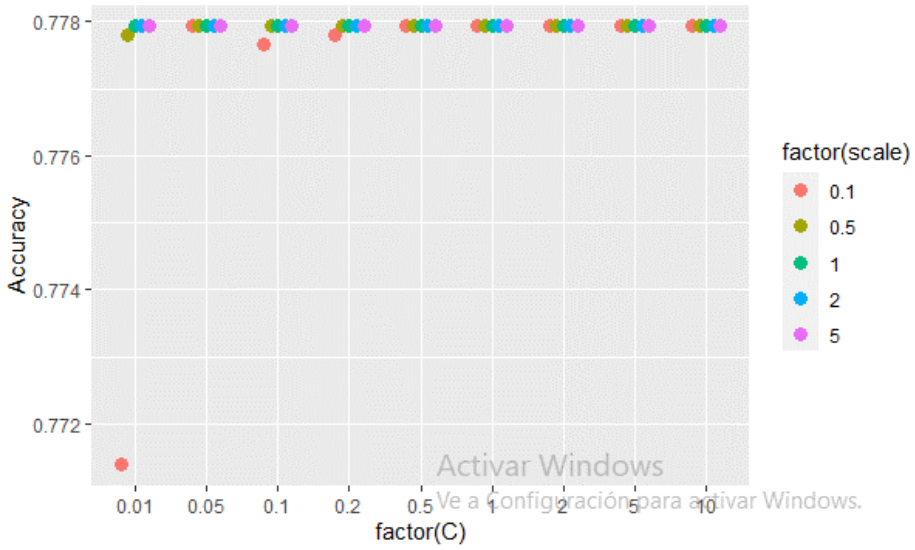


Figura 35 Plot 2 definición de parámetros Scala y C Grupo 1-SVM en R

Grupo	Degree	Scale	C	Accurracy	Kappa	AccurracySD	KappaSD
Grupo 1	2	1	0.01	0.7779375	0.3469185	0.006034642	0.02024904
Grupo 2	3	0.1	0.2	0.7732530	0.3506721	0.012592049	0.024912214
Grupo 3	2	2	0.5	0.7506750	0.235410188	0.0025863421	0.014985855
Grupo 4	3	1	0.2	0.7942642	0.4392560	0.005999159	0.01370688
Grupo 5	3	0.5	0.5	0.7934131	0.43519269	0.005332514	0.02450034

Tabla 39 Resultado de tuneado de SVM polinomial en R

Se realiza el tuneado con la función SVM RBF para cada grupo de variables, al igual que en las otras funciones se tiene en cuenta el parámetro C y Sigma, en el cual prestaremos atención de que al tener un mayor valor de sigma se tendrá menos sesgo pero más varianza, se prueban valores de sigma desde 0,01 a 30, seguido a esto realizamos un plot que nos permite determinar los valores de los factores tuneados, teniendo los siguientes resultados:

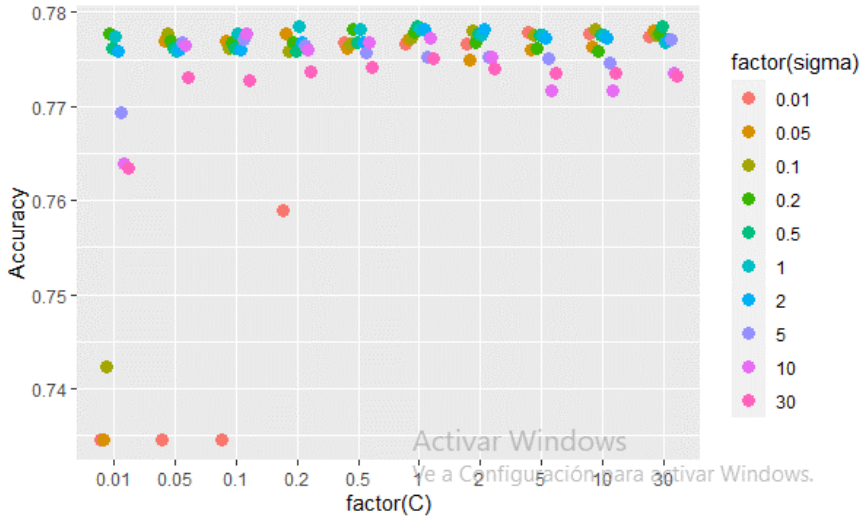


Figura 36 Plot - Definición de parámetros C y Sigma Grupo 1-SVM en R

Grupo	Sigma	C	Accurracy	Kappa
Grupo 1	0.2	0.5	0.7769406	0.33947723
Grupo 2	2	1	0.7786449	0.39496524
Grupo 3	0.5	0.5	0.7418711	0.09673854
Grupo 4	0.5	0.2	0.7951182	0.43505841
Grupo 5	0.01	5	0.7928461	0.40572863

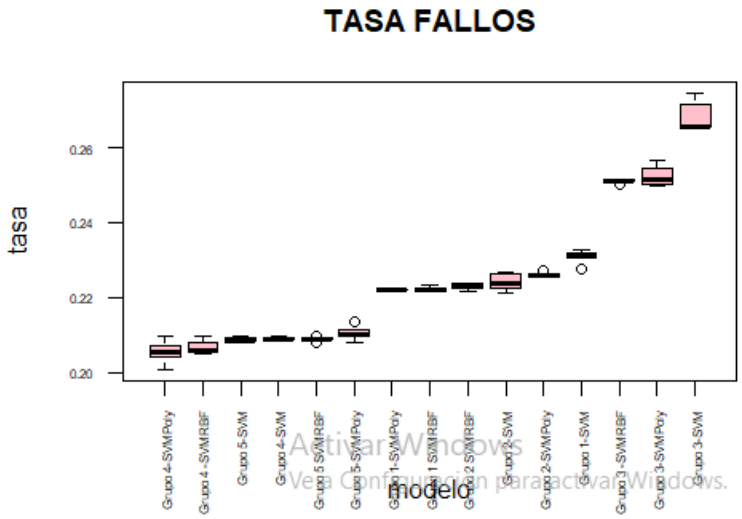
Tabla 40 Resultados de tuneado de SVM RBF en R

Finalizado el tuneado de cada set de variables con las diferentes funciones del kernel, se usan las funciones cruzadassvmbin para el lineal, la función cruzadaSVMbinPoly para el polinomial y cruzadaSVMbinRBF para el RBF, la cuales realizan validación cruzada de 5 repeticiones obteniendo los siguientes resultados:

Grupo	Accurracy	Kappa	AccurracySD	KappaSD
Grupo 1-svm	0.7693044	0.3708059	0.007548468	0.02908505
Grupo 1-svm polinomial	0.7780784	0.3378639	0.007171405	0.02306081
Grupo 1-svm rbf	0.7778229	0.3324621	0.007377895	0.02422716
Grupo 2-svm	0.77611	0.3670027	0.007619276	0.02421706
Grupo 2-svm polinomial	0.7739893	0.3353314	0.006536999	0.01905724
Grupo 2-svm rbf	0.7769424	0.3664765	0.006291028	0.02095729
Grupo 3-svm	0.7338917	0.002301218	0.00224809	0.01299603
Grupo 3-svm polinomial	0.7492263	0.1877014	0.005516947	0.03326111
Grupo 3-svm rbf	0.7487721	0.1853388	0.004970292	0.01370688
Grupo 4-svm	0.7909149	0.4098373	0.007461064	0.02403087
Grupo 4-svm polinomial	0.7946356	0.4211899	0.006827214	0.01954931
Grupo 4-svm rbf	0.7931864	0.4115389	0.007941877	0.02580193
Grupo 5-svm	0.7912838	0.4140357	0.00887361	0.02190683
Grupo 5-svm polinomial	0.7895513	0.3989263	0.007531397	0.02147556
Grupo 5-svm rbf	0.7910566	0.3899969	0.008906131	0.02289518

Tabla 41 Resultados de cruzadas para los diferentes kernel en SVM en R

Se realiza el grafico de los diferentes modelos con cada uno de los sets de variables donde podemos ver que los modelos del set de variables del Grupo 3 presentan mayor variabilidad, una tasa de fallos superior y un AUC más bajo respecto a los otros modelos, los modelos de los sets de variables de los grupos 4 y 5 presentan una tasa de baja y poca variabilidad.



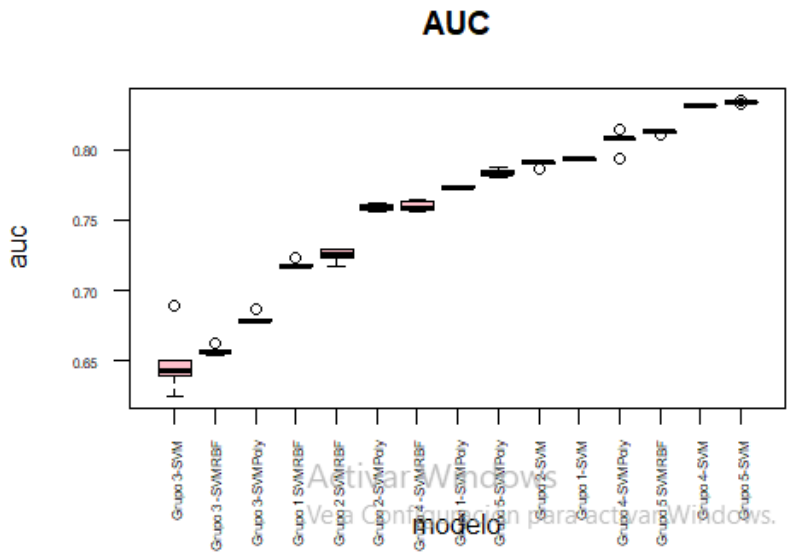


Figura 37 Grafico de cajas de cada set de variables para SVM- tasa de fallos y auc.

Para ver mejor los modelos se toman los que pertenecen a los grupos de variables 4 y 5 ya que son lo que mejores resultados presentan, se toma como ganador el modelo Grupo 5-SVM el cual tiene un kernel lineal, un Accurracy de 0.7912838, una constante C= O.2, presenta poca variabilidad y sesgo.

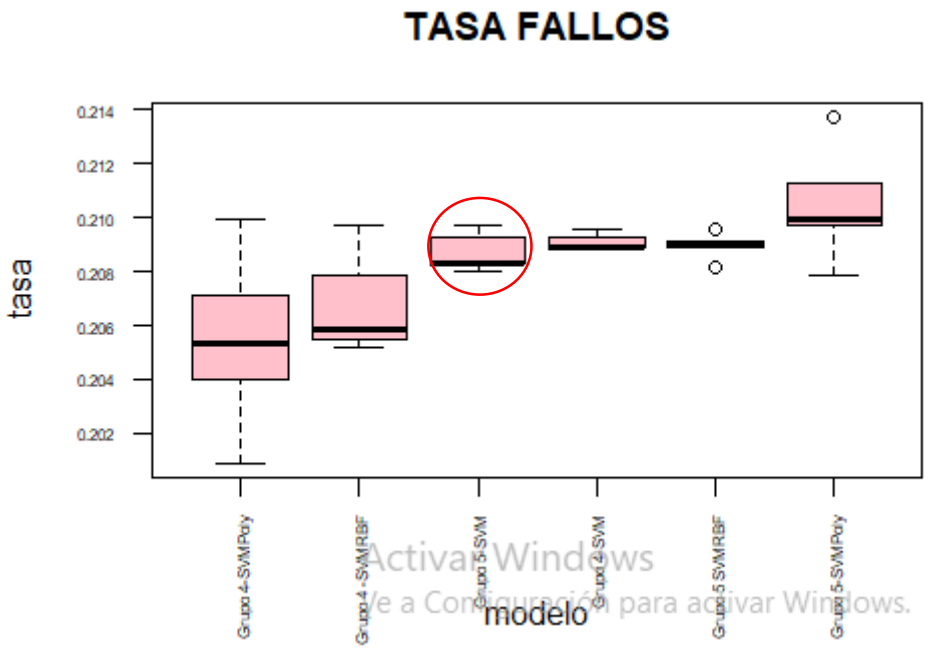


Figura 38 Diagramas de cajas SVM en R - grupos 4 y 5

4.8. Evaluación de modelos

Posterior al tuneado e identificación de los mejores modelos de machine learning, se comparan por medio de validación cruzada y se representan en el mismo grafico los modelos de SAS y R.

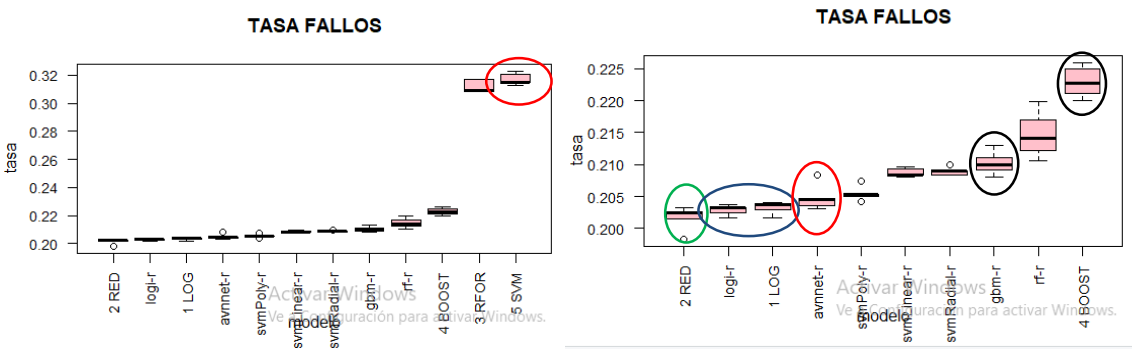


Figura 39 Comparación de modelos en R y SAS

Se observa que los modelos random forest y svm lineal de SAS (grafico de la izquierda) tienen los peores resultados, excluyendo estos (gráfica de la derecha) podemos apreciar mejor a los otros modelos en los cuales se evidencia que la red en SAS (verde) tiene mejores resultados que la regresión de los dos softwares (azul), sin embargo, la red de R (rojo) también obtiene resultados muy cercanos a la regresión logística, es de resaltar que el modelo de gradient boosting de R tiene mejores resultados que el de SAS.

4.9. Ensamblado

Este método nos permite construir predicciones a partir de la combinación de varios modelos, donde se obtienen modelos con un error menor teniendo presente si la correlación entre los clasificadores es menor, menor será el error del ensamblado.

Ensamblado en SAS

Se realiza el ensamblado haciendo uso de la macro crusadastackcon la cual tiene los diferentes algoritmos trabajados en pasos anteriores e incluye sopor vector machine, se realiza la configuración de los diferentes parámetros de los modelos seleccionados como ganadores, obteniendo las predicciones de cada algoritmo, para esto se realiza validación cruzada con 5 repeticiones, en total se construyen 23 modelos con los ensamblados.

Ensamblado en R

Se realiza el ensamblado de los algoritmos anteriormente trabajados, se usa el archivo cruzadaensambladobinaria la cual contiene todas las funciones anteriormente trabajadas en cada algoritmo y se compilan para el ensamblado, posterior a esto se realiza validación cruzada repetida para cada algoritmo, realizando 42 modelos con ensamblados.

En el siguiente diagrama de cajas se representan los algoritmos individuales (en verde SAS y en azul R) y los ensamblados de R tiene prefijo pred los restantes son de SAS.

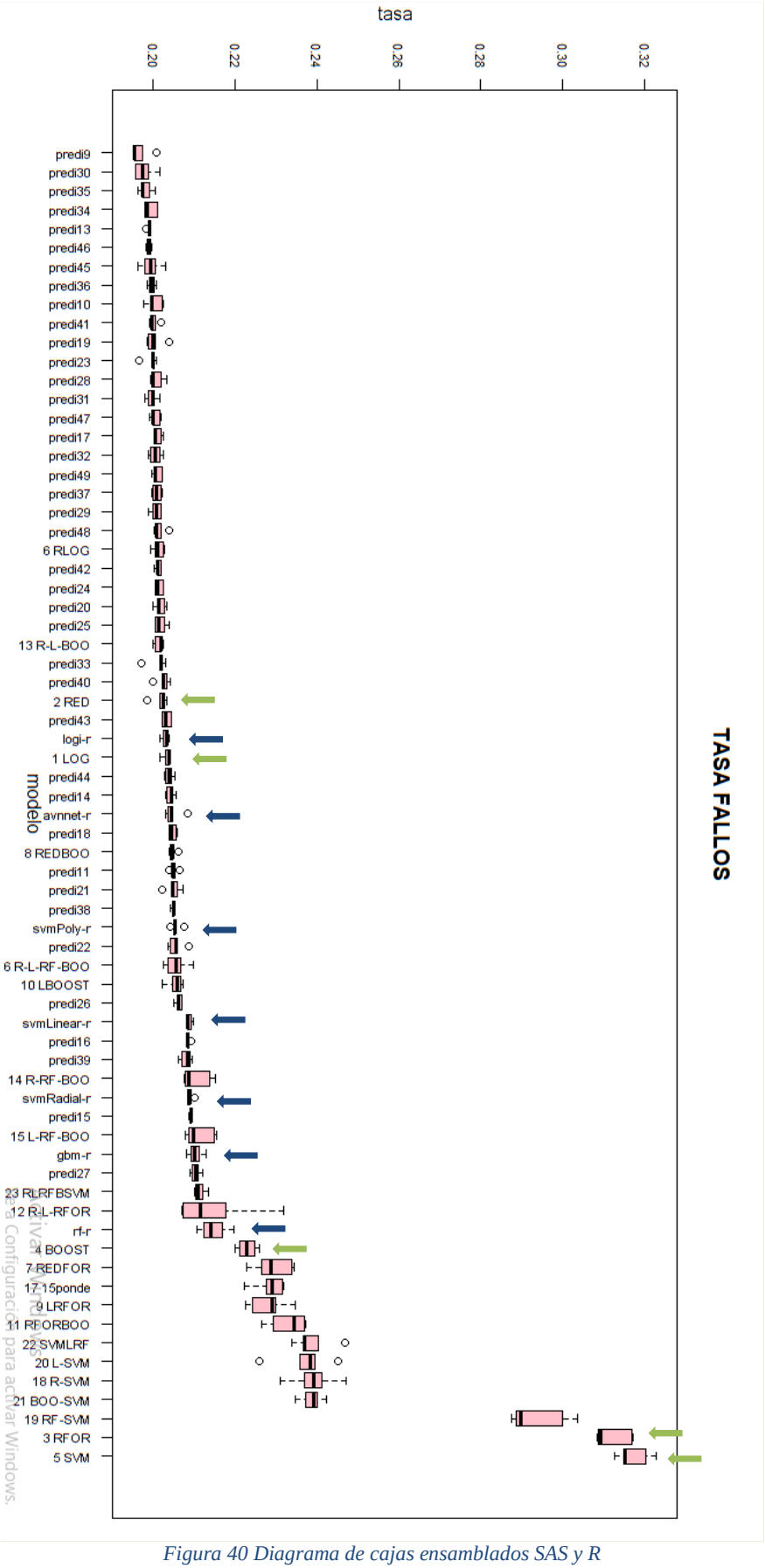


Figura 40 Diagrama de cajas ensamblados SAS y R

Se observa como random forest y svm lineal en SAS presenta una tasa de fallos alta al igual que su variabilidad, en cuanto a los ensamblados de estos modelos vemos como disminuye el error como es el caso del ensamblado redfor de SAS que corresponde a la red que presenta una tasa de error baja con respecto a random forest, pero a pesar de que esta mejora se aumenta la variabilidad del modelo.

Los modelos de regresión logística de ambos algoritmos son superados por 31 ensamblados, donde 2 corresponden a SAS (ensamblado 1: red, logística, ensamblado 2: red, logística y boosting) y el restante corresponden a R, lo que nos permite determinar que los modelos realizados en R tienen un mejor desempeño de acuerdo al tuneado de sus parámetros.

Para ver más el detalle y poder seleccionar el modelo ganador se seleccionan los siguientes ensamblados y se deja de referencia los modelos principales:

Descripcion de ensamblados	
Predi46 (random forest + avnnet + svmradial)	Predi44 (random forest +gbm + svmradial)
Predi36 (logistica+ forest random + svmradial)	Predi26 (gbm + svmpoly)
Predi24 (random forest +svmradial)	Predi27 (gbm + svmradial)
Predi42 (random forest +gbm + svmradial)	

Tabla 42 Descripción de modelos seleccionados en ensamblado

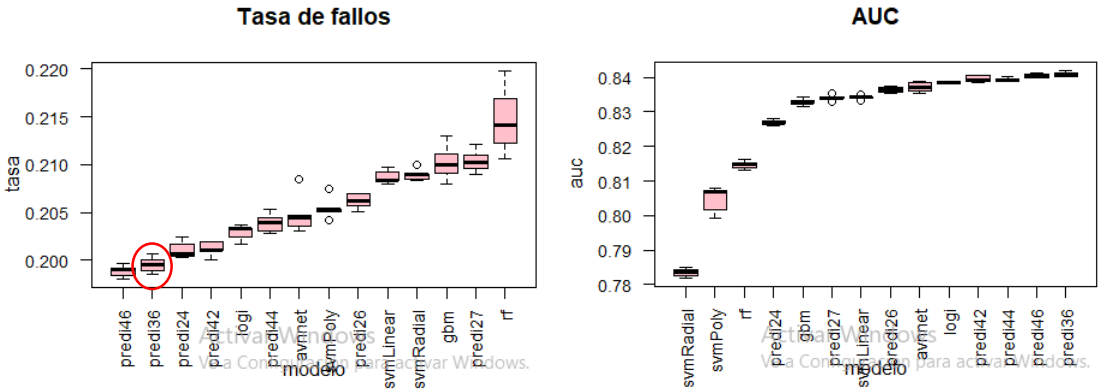


Figura 41 Diagrama de cajas modelos seleccionados en ensamblado

Se puede observar que el ensamblado predi46 que cuenta con tres algoritmos, reduce bastante la varianza y sesgo respecto a sus algoritmos originales, al igual que el modelo predi 36, el cual es bastante simétrico frente a la regresión logística, teniendo en cuenta esto se toma como ganador el modelo predi36 con ensamblado de logística, random forest y svmradial con una tasa de fallos promedio de 0,19952 y un AUC 0,84090, siendo los resultados de la regresión logística 0,20279 y 0,83846 respectivamente, presentado una mejora muy pequeña en la tasa de fallos y un incremento leve en el AUC.

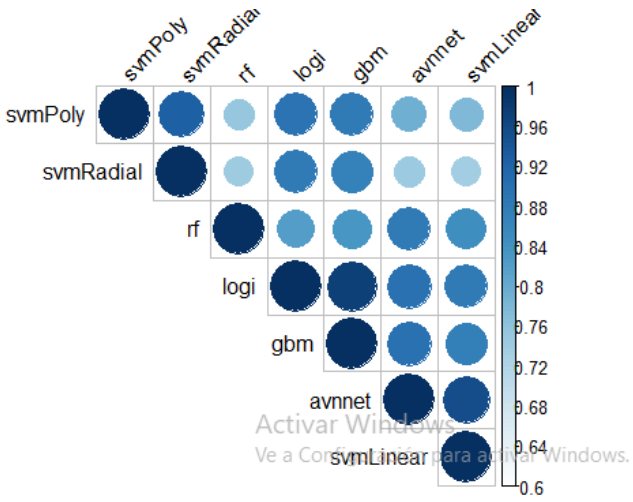


Figura 42 Correlación entre predicciones ensamblado en R

Al observar el grafico de correlaciones de los algoritmos principales de R podemos ver que la correlación más fuerte esta en logística y gradient boosting al igual que en la red y svmlineal. Pero al ver la correlación entre savmradial y random forest es baja esto permite reducir el sesgo en el ensamble.

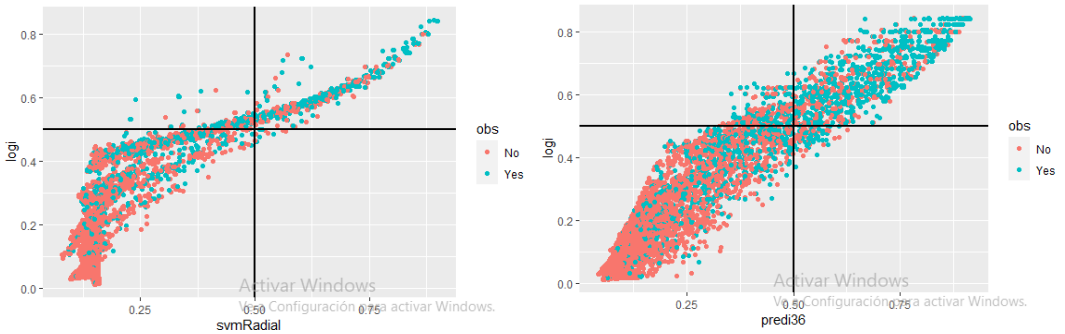


Figura 43 Correlación de predicciones svmradial y logística / predi 36 y logística R

En el cuadrante inferior izquierdo de la correlación entre svmradial y logística vemos que svmradial considera que los puntos en esta zona son rojos que corresponden a la logística y falla en los verdes, sin embargo, al ver la diagonal de izquierda a derecha vemos que cuenta con una gran cantidad de puntos en los que los algoritmos están de acuerdo.

En cuanto a la correlación entre el modelo predi36 y la logística vemos que la distribución de los puntos tiene un rango más amplio, siendo mucho más claro las zonas donde predi36 indica el cuadrante inferior izquierdo que los puntos en color rojo pertenecen a la logística y los verdes son una falla siendo pocos.

5. Conclusiones

El objetivo principal de esta investigación es poder encontrar el modelo que prediga los clientes que pueden cancelar la suscripción en una empresa de telecomunicaciones, esta medición es realizada por medio del indicador Churn (tasa de cancelación) , enfocándonos en el Churn voluntario, los modelos se han realizado con una muestra de 7043 registros y 21 variables incluyendo la variable objetivo Churn, las cuales contenían información sobre el cliente, los servicios contratados e información demográfica, se aplicaron diferentes métodos de selección de variables , se formaron 5 grupos de variables que fueron probadas con los diferentes algoritmos de machine learning y de acuerdo a los resultados de los modelos poder identificar el mejor set de variables que más aportara a la variable objetivo, teniendo en cuenta lo anterior la mayoría los modelos seleccionados como ganadores de cada algoritmo contenían el set de variables del grupo 4 el cual tiene las siguientes variables: Antigüedad del cliente, tipo de contrato mes a mes, cargos totales, clientes con seguridad en línea, servicio de internet con fibra óptica y facturación electrónica.

Se realizó la investigación con 6 algoritmos de machine learning: regresión lineal, redes neuronales, bagging, random forest, gradient boosting y support vector machine con kernel lineal, polinomial y RDF, se tomaron las mejores versiones de cada modelo y mediante el método de ensamblado se obtuvieron 65 modelos incluyendo los principales, el ensamblado fue realizado en cada software (23 en SAS y 42 en R), posteriormente en una gráfica se unificaron sus diagramas de cajas seleccionado como ganador el ensamblado de los modelos regresión logística, random forest y svmradial del software R, con una tasa de fallos de 0,19952 y un AUC 0,84090, sin embargo al compararlo con un modelo tradicional como la regresión logística con una tasa de fallos de 0,20279, se evidencia una mejora pequeña lo cual nos llevaría a seleccionar este modelo como ganador teniendo en cuenta que es menos complejo.

Al tener la oportunidad de trabajar los algoritmos en dos softwares SAS y R se puede evidenciar que el tuneado de los parámetros en R representan una mejora en los resultados de los modelos, evidenciándose no solo en los modelos principales si no especialmente en los ensamblados donde 29 de los 32 que se encontraban con mejores resultados respecto a la regresión logística son de R y tan solo 2 modelos provenían de SAS.

Como oportunidades de mejora se considera que la inclusión de variables relacionadas con incidencias técnicas o problemas de facturación podrían aportarle bastante al modelo ya que el set de variables actual no las contiene, en cuanto futuras líneas de trabajo se complementaría el análisis predictivo con la generación de un dashboard que permita tener una visual rápida y clara de los clientes potencial a cancelar el servicio, facilitando la generación de estrategias de fidelización y la elaboración de campañas proactivas que le permitan a la empresa disminuir el Churn y mejorar su relación con el cliente.

6. Bibliografía

[1]-Reche, C. EDeconomiaDigital (2019, Mayo 17). Estas son la operadoras que tienen los clientes más infieles, https://www.economiadigital.es/directivos-y-empresas/orange-vodafone-masmovil-clientes-mas-infieles_625325_102.html.

[2]- landa, J. (2016, Febrero 19). Tratamiento de los datos, de http://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/.

[3]- Revista Ingeniería de Sistemas, Volumen XXVII. (2013, Septiembre). Aplicación de Minería de Datos para Predecir Fuga de Clientes En La Industria De Las Telecomunicaciones, de https://silo.tips/download/aplicacion-de-mineria-de-datos-para-predecir-fuga-de-clientes-en-la-industria-de.

[4]- Andrade, F. (2014., Aproximación a los factores determinantes del Churn desde un enfoque de Marketing Relacional innovador: el punto de vista de los proveedores y clientes de Servicios, de https://repositorio.uam.es/bitstream/handle/10486/660735/andrade_fernanda_maria_de_%20jesus.pdf?sequence=1&isAllowed=y.

[5]- Cesari, J. (2018). Modelo predictivo para determinar la Tasa Churn en pacientes de un centro médico, de https://www.mti.cl/wp-content/uploads/2018/12/Tesina_2018_Cesari-Jean.pdf.

[6]- Mesa, A. (2018). Predicción de fuga de clientes en una empresa de telefonía utilizando el algoritmo adaboost desbalanceado y la regresión logística asimétrica, de http://repositorio.lamolina.edu.pe/bitstream/handle/UNALM/3245/meza-rodriguez-aldo-richard.pdf?sequence=1&isAllowed=y.

[7]- Deroche, A., Basso, D - Pollo., M (2019, julio). Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza, Aplicación de algoritmos de aprendizaje automático al análisis del churn en planes de ahorro, de https://reddi.unlam.edu.ar/index.php/ReDDi/article/download/82/164?inline=1.

[8]- González, L. (2019, Junio 28). Regresión logistica-teoria, de https://ligdigonzalez.com/regresion-logistica-multiple-machine-learning-teoria/.

[9]- Calvo, D. (2017, Julio 12). Definición de red neuronal artificial, de https://www.diegocalvo.es/definicion-de-red-neuronal/#:~:text=autom%C3%A1tico%20%7C%201%20Comentario-,Definici%C3%B3n%20de%20red%20neuronal%20artificial,interconectadas%20entre%20s%C3%AD%20mediante%20enlaces.

[10]- Santa, E. (2014, Diciembre 14). Bagging para mejorar un modelo predictivo, de http://apuntes-r.blogspot.com/2014/12/bagging-para-mejorar-un-modelo.html.

[11]- Portela, J. (2019). Apuntes de la signatura Machine Learning.

[12]- Mathworks. (2020). Máquinas de vectores de soporte, de https://la.mathworks.com/discovery/support-vector-machine.html.

[13]- Sancho, F. (2018, Diciembre 26). Métodos combinados de aprendizaje, de http://www.cs.us.es/~fsancho/?e=106.

7. Anexos

Randomselectlog

```
libname red 'C:\Users\michelle\Documents\TFM CHURN\disco c';
data uno; set red.em_save_train;


%randomselectlog(data=uno,
listclass=OPT_IMP_REP_TotalCharges OPT_REP_MonthlyCharges,
vardepen=Churn,
modelo= TI_Partner1 TI_PaperlessBilling2 TI_Partner2
TI_OnlineSecurity3 TI_PaperlessBilling1
        TI_PaymentMethod3 TI_PaymentMethod4 TI_PaymentMethod1
TI_PaymentMethod2 TI_MultipleLines3
        TI_OnlineBackup1 TI_MultipleLines1 TI_MultipleLines2
TI_OnlineSecurity1 TI_OnlineSecurity2
        TI_OnlineBackup2 TI_OnlineBackup3 TI_TechSupport1
TI_TechSupport2 TI_StreamingTV2
        TI_StreamingTV3 TI_gender2 numMissing TI_TechSupport3
TI_gender1 TI_SeniorCitizen1
        TI_SeniorCitizen2 TI_PhoneService1 TI_PhoneService2
TI_StreamingMovies3 TI_StreamingTV1
        TI_StreamingMovies1 TI_StreamingMovies2 TI_Contract2
TI_Contract1 TI_Dependents1 TI_Contract3
        LOG_REP_tenure EXP_var_aleatoria OPT_IMP_REP_TotalCharges
OPT_REP_MonthlyCharges
        TI_InternetService1 TI_DeviceProtection3 TI_InternetService3
TI_InternetService2
        TI_DeviceProtection1 TI_Dependents2 TI_DeviceProtection2,
sinicio=12345,
sfinal=12400,
fracciontrain=0.8,
directorio=C:\Users\michelle\Documents\TFM CHURN\disco c);
```

Proc logistic

```
proc logistic data=uno namelen=20 descending ;
class OPT_IMP_REP_TotalCharges OPT_REP_MonthlyCharges ;
model Churn=TI_Partner1 TI_PaperlessBilling2 TI_Partner2
TI_OnlineSecurity3 TI_PaperlessBilling1
        TI_PaymentMethod3 TI_PaymentMethod4 TI_PaymentMethod1
TI_PaymentMethod2 TI_MultipleLines3
        TI_OnlineBackup1 TI_MultipleLines1 TI_MultipleLines2
TI_OnlineSecurity1 TI_OnlineSecurity2
        TI_OnlineBackup2 TI_OnlineBackup3 TI_TechSupport1
TI_TechSupport2 TI_StreamingTV2
        TI_StreamingTV3 TI_gender2 numMissing TI_TechSupport3
TI_gender1 TI_SeniorCitizen1
        TI_SeniorCitizen2 TI_PhoneService1 TI_PhoneService2
TI_StreamingMovies3 TI_StreamingTV1
        TI_StreamingMovies1 TI_StreamingMovies2 TI_Contract2
TI_Contract1 TI_Dependents1 TI_Contract3
        LOG_REP_tenure EXP_var_aleatoria OPT_IMP_REP_TotalCharges
OPT_REP_MonthlyCharges
        TI_InternetService1 TI_DeviceProtection3 TI_InternetService3
TI_InternetService2
        TI_DeviceProtection1 TI_Dependents2 TI_DeviceProtection2
/selection=stepwise;
```

```
run;quit;
data mode;length effect $20. modelo $ 20000;retain modelo " ";set
parametros end=fin;effect=cat(' ',effect);
if _n_ ne 1 then modelo=catt(modelo,' ',effect);if fin then output;
run;
data ;set mode;put modelo;run;
```

macro variar pruebas con la red

```
/*GRUPO 1* ALGORITMO LEVMARK*/
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=LOG_REP_tenure,
   listclass=TI_Contract1
TI_TechSupport1,vardep=Churn,nodos=&nodos,corte=50,semilla=&semilla,po
rcen=0.80,algo=levmar);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;
%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=70,incr
enodos=6);

/*GRUPO 2* ALGORITMO LEVMARK*/
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=OPT_IMP_REP_TotalCharges,
   listclass=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,vardep=Churn,nodos=&nodos,corte=50,semilla=&semill
a,porcen=0.80,algo=levmar);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;
%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=39,incr
enodos=6);


/*GRUPO 3* ALGORITMO LEVMARK*/
```

```
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=OPT_REP_MonthlyCharges,
   listclass=TI_InternetService2 TI_PaperlessBilling1
TI_SeniorCitizen1,vardep=Churn,nodos=&nodos,corte=50,semilla=&semilla,
porcen=0.80,algo=levmar);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;
%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=58,incr
enodos=6);


/*GRUPO 4* ALGORITMO LEVMARK*/
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
   listclass=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,vardep=Churn,nodos=&nodos,corte=50,semilla=&semil
la,porcen=0.80,algo=levmar);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=44,incr
enodos=6);


/*GRUPO 5* ALGORITMO LEVMARK*/
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
   listclass=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
```

```
TI_StreamingMovies3,vardep=Churn,nodos=&nodos,corte=50,semilla=&semill
a,porcen=0.80,algo=levmar);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=29,incr
enodos=6);


/* IGUAL CON BPROP */

/*Grupo 1*/
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=LOG_REP_tenure,
   listclass=TI_Contract1
TI_TechSupport1,vardep=Churn,nodos=&nodos,corte=50,semilla=&semilla,po
rcen=0.80,algo=bprop mom=0.2 learn=0.1);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=70,incr
enodos=6);


/*Grupo 2* Bprop*/
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=OPT_IMP_REP_TotalCharges,
   listclass=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,vardep=Churn,nodos=&nodos,corte=50,semilla=&semill
a,porcen=0.80,algo=bprop mom=0.2 learn=0.1);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
```

```
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=39,incr
enodos=6);


/*Grupo 3 Bprop*/
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=OPT_REP_MonthlyCharges,
   listclass=TI_InternetService2 TI_PaperlessBilling1
TI_SeniorCitizen1,vardep=Churn,nodos=&nodos,corte=50,semilla=&semilla,
porcen=0.80,algo=bprop mom=0.2 learn=0.1);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=58,incr
enodos=6);

/*Grupo 4 Bprop*/
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
   listclass=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,vardep=Churn,nodos=&nodos,corte=50,semilla=&semil
la,porcen=0.80,algo=bprop mom=0.2 learn=0.1);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=44,incr
enodos=6);

/*Grupo 5 Bprop*/
```

```
%macro
variar(seminicio=,semifin=,inicionodos=,finalnodos=,increnodos=);
title '';
data union;run;
%do semilla=&seminicio %to &semifin;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
   %neuralbinariabasica(archivo=uno,
   listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
   listclass=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,vardep=Churn,nodos=&nodos,corte=50,semilla=&semill
a,porcen=0.80,algo=bprop mom=0.2 learn=0.1);
   data estadisticos;set
estadisticos;nodos=&nodos;semilla=&semilla;run;
   data union;set union estadisticos;run;
%end;
%end;
proc sort data=union;by nodos;run;
proc boxplot data=union;plot (porcenVN porcenFN porcenVP porcenFP
sensi especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%variar(seminicio=12345,semifin=12348,inicionodos=1,finalnodos=29,incr
enodos=6);

/*EARLY STOPPING ?*/

/* Grupo 1*/
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=13,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=19,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=13,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=19,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=13,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=19,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
```

```
vardep=Churn,porcen=0.80,semilla=442712,ocultos=13,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=19,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=13,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=19,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=13,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=19,meto=levmar,acti=TA
NH);



/* Grupo 2*/
%redneuronalbinaria(archivo=uno,listclass=TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listconti=OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=13,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listconti=OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=9,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=13,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=9,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listconti=OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=13,meto=bprop,acti=TAN
H);
```

```
%redneuronalbinaria(archivo=uno,listclass=TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listconti=OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=9,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=13,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=9,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listconti=OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=13,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listconti=OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=9,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=13,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_TechSupport1,
listconti=LOG_REP_tenure,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=9,meto=levmar,acti=TAN
H);


/* Grupo 3*/
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=17,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=17,meto=levmar,acti=TA
NH);
```

```
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=17,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=17,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=17,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listconti=OPT_REP_MonthlyCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=17,meto=levmar,acti=TA
NH);


/* Grupo 4*/
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=10,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
```

```
vardep=Churn,porcen=0.80,semilla=442711,ocultos=10,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=10,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=10,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=10,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=10,meto=levmar,acti=TA
NH);


/* Grupo 5*/
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=10,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
```

```
vardep=Churn,porcen=0.80,semilla=442711,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442711,ocultos=10,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=10,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442712,ocultos=10,meto=levmar,acti=TA
NH);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=7,meto=bprop,acti=TANH
);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=10,meto=bprop,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=7,meto=levmar,acti=TAN
H);
%redneuronalbinaria(archivo=uno,listclass=TI_Contract1 TI_TechSupport1
TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1
TI_Contract2 TI_StreamingMovies3,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
vardep=Churn,porcen=0.80,semilla=442713,ocultos=10,meto=levmar,acti=TA
NH);
```

Baggin sas

```
/*Bagging Grupo 1*/
```

```
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,
maxtrees=100,variables=3,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final1;set final;modelo=1;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,
maxtrees=100,variables=3,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final2;set final;modelo=2;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,
maxtrees=100,variables=3,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final3;set final;modelo=3;

data union;set final1 final2 final3 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,
maxtrees=300,variables=3,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.05,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final4;set final;modelo=4;

data union;set final1 final2 final3 final4;
proc boxplot data=union;plot media*modelo;run;


/*Bagging Grupo 2*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
maxtrees=100,variables=7,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final5;set final;modelo=5;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
```

```
maxtrees=100,variables=7,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final6;set final;modelo=6;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
maxtrees=100,variables=7,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final7;set final;modelo=7;

data union;set final5 final6 final7 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
maxtrees=100,variables=7,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.05,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final8;set final;modelo=8;

data union;set final5 final6 final7 final8;
proc boxplot data=union;plot media*modelo;run;


/*Bagging Grupo 3*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1 TI_SeniorCitizen1,
maxtrees=100,variables=4,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final9;set final;modelo=9;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1 TI_SeniorCitizen1,
maxtrees=100,variables=4,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final10;set final;modelo=10;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1 TI_SeniorCitizen1,
maxtrees=100,variables=4,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final11;set final;modelo=11;
```

```
data union;set final9 final10 final11 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1 TI_SeniorCitizen1,
maxtrees=100,variables=4,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.05,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final12;set final;modelo=12;

data union;set final9 final10 final11 final12;
proc boxplot data=union;plot media*modelo;run;


/*Bagging Grupo 4*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,
maxtrees=100,variables=6,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final13;set final;modelo=13;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,
maxtrees=100,variables=6,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final14;set final;modelo=14;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,
maxtrees=100,variables=6,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final15;set final;modelo=15;

data union;set final13 final14 final15 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,
maxtrees=100,variables=6,porcenbag=0.80,maxbranch=4,tamhoja=20,maxdept
h=5,pvalor=0.01,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final16;set final;modelo=16;
```

```
data union;set final13 final14 final15 final16;
proc boxplot data=union;plot media*modelo;run;


/*Bagging Grupo 5*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
maxtrees=100,variables=10,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdept
h=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final17;set final;modelo=17;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
maxtrees=100,variables=10,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdep
th=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final18;set final;modelo=18;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
maxtrees=100,variables=10,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdep
th=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final19;set final;modelo=19;

data union;set final17 final18 final19 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
maxtrees=200,variables=10,porcenbag=0.80,maxbranch=4,tamhoja=8,maxdept
h=3,pvalor=0.05,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final20;set final;modelo=20;

data union;set final17 final18 final19 final20;
proc boxplot data=union;plot media*modelo;run;

/*grafica de todos los modelos de baggin*/
data union;set final1 final2 final3 final4 final5 final6 final7 final8
final9 final10 final11 final12 final13 final14 final15 final16 final17
final18 final19 final20;
```

```
proc boxplot data=union;plot media*modelo;run;

/*grafica de los modelos ganadores por set de variables*/

data union;set final2 final8 final10  final15  final19 ;
proc boxplot data=union;plot media*modelo;run;
```

Random forest

```
/*forest Grupo 1*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,
maxtrees=100,variables=2,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final21;set final;modelo=21;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,
maxtrees=100,variables=2,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final22;set final;modelo=22;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,
maxtrees=100,variables=2,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final23;set final;modelo=23;

data union;set final21 final22 final23 ;
proc boxplot data=union;plot media*modelo;run;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,
maxtrees=100,variables=2,porcenbag=0.80,maxbranch=4,tamhoja=20,maxdept
h=4,pvalor=0.05,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final24;set final;modelo=24;

data union;set final21 final22 final23 final24;
proc boxplot data=union;plot media*modelo;run;


/*Random forest Grupo 2*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
```

```
maxtrees=100,variables=6,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final25;set final;modelo=25;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
maxtrees=100,variables=5,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final26;set final;modelo=26;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
maxtrees=100,variables=6,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final27;set final;modelo=27;

data union;set final25 final26 final27 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
maxtrees=200,variables=5,porcenbag=0.80,maxbranch=4,tamhoja=10,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final28;set final;modelo=28;

data union;set final25 final26 final27 final28;
proc boxplot data=union;plot media*modelo;run;


/*ramdom forest Grupo 3*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1 TI_SeniorCitizen1,
maxtrees=100,variables=3,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final29;set final;modelo=29;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1 TI_SeniorCitizen1,
maxtrees=100,variables=3,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final30;set final;modelo=30;
```

```
%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1 TI_SeniorCitizen1,
maxtrees=100,variables=2,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final31;set final;modelo=31;

data union;set final29 final30 final31 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1 TI_SeniorCitizen1,
maxtrees=100,variables=3,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.05,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final32;set final;modelo=32;

data union;set final29 final30 final31 final32;
proc boxplot data=union;plot media*modelo;run;


/*Random Grupo 4*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,
maxtrees=100,variables=5,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final33;set final;modelo=33;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,
maxtrees=100,variables=5,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final34;set final;modelo=34;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,
maxtrees=100,variables=4,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final35;set final;modelo=35;

data union;set final33 final34 final35 ;
proc boxplot data=union;plot media*modelo;run;
```

```
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1 TI_InternetService2
TI_PaperlessBilling1,
maxtrees=100,variables=4,porcenbag=0.80,maxbranch=4,tamhoja=30,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final36;set final;modelo=36;

data union;set final33 final34 final35 final36 ;
proc boxplot data=union;plot media*modelo;run;


/*Bagging Grupo 5*/
%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
maxtrees=100,variables=9,porcenbag=0.80,maxbranch=4,tamhoja=5,maxdepth
=10,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final37;set final;modelo=37;

%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
maxtrees=100,variables=8,porcenbag=0.80,maxbranch=4,tamhoja=15,maxdept
h=4,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final38;set final;modelo=38;

%cruzadarandomforestbin(
archivo=Uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
maxtrees=100,variables=8,porcenbag=0.80,maxbranch=4,tamhoja=25,maxdept
h=5,pvalor=0.1,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final39;set final;modelo=39;

data union;set final37 final38 final39 ;
proc boxplot data=union;plot media*modelo;run;


%cruzadarandomforestbin(
archivo=uno,vardep=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
maxtrees=100,variables=8,porcenbag=0.80,maxbranch=4,tamhoja=20,maxdept
h=4,pvalor=0.05,
```

```
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final40;set final;modelo=40;

data union;set final37 final38 final39 final40;
proc boxplot data=union;plot media*modelo;run;


/* GRAFICA MODELO GANADOR*/

data union;set final24 final25 final30 final34 final40;
proc boxplot data=union;plot media*modelo;run;

data union;set  final25 final30 final34 final40;
proc boxplot data=union;plot media*modelo;run;
```

Gradient boosting sas

```
/*Grupo 1 gradient boosting*/
%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,leafsize=15,
iteraciones=300,shrink=0.05,maxbranch=4,maxdepth=4,mincatsize=15,minob
s=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final41;set final;modelo=41;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,leafsize=10,
iteraciones=200,shrink=0.1,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final42;set final;modelo=42;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=LOG_REP_tenure,
categor=TI_Contract1 TI_TechSupport1,leafsize=15,
iteraciones=300,shrink=0.2,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final43;set final;modelo=43;


data union;set final41 final42 final43 ;
proc boxplot data=union;plot media*modelo;run;


/*Grupo 2 gradient boosting*/

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,leafsize=15,
iteraciones=300,shrink=0.05,maxbranch=4,maxdepth=4,mincatsize=15,minob
s=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final44;set final;modelo=44;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=OPT_IMP_REP_TotalCharges,
```

```
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,leafsize=10,
iteraciones=200,shrink=0.1,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final45;set final;modelo=45;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=OPT_IMP_REP_TotalCharges,
categor=TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,leafsize=15,
iteraciones=300,shrink=0.2,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final46;set final;modelo=46;



data union;set final44 final45 final46 ;
proc boxplot data=union;plot media*modelo;run;


/*Grupo 3 gradient boosting*/

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1
TI_SeniorCitizen1,leafsize=15,
iteraciones=300,shrink=0.05,maxbranch=4,maxdepth=4,mincatsize=15,minob
s=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final49;set final;modelo=49;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1
TI_SeniorCitizen1,leafsize=10,
iteraciones=200,shrink=0.1,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final50;set final;modelo=50;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=OPT_REP_MonthlyCharges,
categor=TI_InternetService2 TI_PaperlessBilling1
TI_SeniorCitizen1,leafsize=15,
iteraciones=300,shrink=0.2,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final51;set final;modelo=51;



data union;set final49 final50 final51 ;
proc boxplot data=union;plot media*modelo;run;



/*Grupo 4 gradient boosting*/

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
```

```
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1
TI_InternetService2,leafsize=15,
iteraciones=300,shrink=0.05,maxbranch=4,maxdepth=4,mincatsize=15,minob
s=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final52;set final;modelo=52;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1
TI_InternetService2,leafsize=10,
iteraciones=200,shrink=0.1,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final53;set final;modelo=53;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_OnlineSecurity1
TI_InternetService2,leafsize=15,
iteraciones=300,shrink=0.2,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final54;set final;modelo=54;



data union;set final52 final53 final54 ;
proc boxplot data=union;plot media*modelo;run;


/*Grupo 5 gradient boosting*/

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,leafsize=15,
iteraciones=300,shrink=0.05,maxbranch=4,maxdepth=4,mincatsize=15,minob
s=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final55;set final;modelo=55;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,leafsize=10,
iteraciones=200,shrink=0.1,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final56;set final;modelo=56;

%cruzadatreeboostbin(archivo=uno,vardepen=Churn,
conti=LOG_REP_tenure OPT_IMP_REP_TotalCharges,
categor=TI_Contract1 TI_TechSupport1 TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,leafsize=15,
iteraciones=300,shrink=0.2,maxbranch=4,maxdepth=4,mincatsize=15,minobs
=20,
```

```
ngrupos=4,sinicio=13345,sfinal=13349,objetivo=tasafallos);
data final57;set final;modelo=57;


data union;set final55 final56 final57 ;
proc boxplot data=union;plot media*modelo;run;

/*modelos ganadores por set de variables*/
data union;set final43 final44 final50 final52 final55 ;
proc boxplot data=union;plot media*modelo;run;

/*modelos ganadores por set de variables omitinedo el 52*/
data union;set final43 final44 final52 final55 ;
proc boxplot data=union;plot media*modelo;run;
```

SVM EN SAS

```
/*Grupo 1*/
%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure TI_Contract1 TI_TechSupport1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=10,c=10);
data final59;set final;modelo='Grupo1-59-SVM-lineal';

%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure TI_Contract1 TI_TechSupport1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=2,c=10);
data final60;set final;modelo='Grupo1-60-SVM-lineal';

%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure TI_Contract1 TI_TechSupport1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=3,c=10);
data final61;set final;modelo='Grupo1-61-SVM-lineal';

/*grupo 2*/

%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=OPT_IMP_REP_TotalCharges TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=10,c=10);
data final62;set final;modelo='Grupo2-62-SVM-lineal';


%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=OPT_IMP_REP_TotalCharges TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=2,c=10);
data final63;set final;modelo='Grupo2-63-SVM-lineal';
```

```
%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=OPT_IMP_REP_TotalCharges TI_OnlineSecurity1
TI_PaymentMethod3 TI_StreamingTV3 TI_MultipleLines1 TI_Contract2
TI_StreamingMovies3,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=3,c=10);
data final64;set final;modelo='Grupo2-64-SVM-lineal';

/*grupo 3*/
%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=OPT_REP_MonthlyCharges TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=10,c=10);
data final65;set final;modelo='Grupo3-65-SVM-lineal';


%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=OPT_REP_MonthlyCharges TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=2,c=10);
data final66;set final;modelo='Grupo3-66-SVM-lineal';

%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=OPT_REP_MonthlyCharges TI_InternetService2
TI_PaperlessBilling1 TI_SeniorCitizen1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=3,c=10);
data final67;set final;modelo='Grupo3-67-SVM-lineal';


/*grupo 4*/
%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=10,c=10);
data final68;set final;modelo='Grupo4-68-SVM-lineal';


%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=2,c=10);
data final69;set final;modelo='Grupo4-69-SVM-lineal';

%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges TI_Contract1
TI_OnlineSecurity1 TI_InternetService2 TI_PaperlessBilling1,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=3,c=10);
data final70;set final;modelo='Grupo4-70-SVM-lineal';
```

```
/*grupo 5*/
%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges TI_Contract1
TI_TechSupport1 TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=10,c=10);
data final71;set final;modelo='Grupo5-71-SVM-lineal';


%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges TI_Contract1
TI_TechSupport1 TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=2,c=10);
data final72;set final;modelo='Grupo5-72-SVM-lineal';

%cruzadaSVMbin
(archivo=Sovm,vardepen=Churn,
listconti=LOG_REP_tenure OPT_IMP_REP_TotalCharges TI_Contract1
TI_TechSupport1 TI_OnlineSecurity1 TI_PaymentMethod3 TI_StreamingTV3
TI_MultipleLines1 TI_Contract2 TI_StreamingMovies3,
listclass=,
ngrupos=4,sinicio=12345,sfinal=12349,kernel=linear k_par=3,c=10);
data final73;set final;modelo='Grupo5-73-SVM-lineal';

options notes;

data union;length modelo $ 40;set final59 final60 final61 final62
final63 final66 final65 final66 final67 final68 final69 final70
final71 final72 final73;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;

/*MODELOS GANADORES*/

data union;length modelo $ 40;set final60 final63 final66 final69
final72;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;


REDES EN R

# Grupo 1 redes
medias2<-cruzadaavnnetbin(data=em_save_TRAINbis,

vardep="Churn",listconti=c("LOG_REP_tenure","TI_Contract1","TI_TechSup
port1"),
  listclass=c(""),grupos=4,sinicio=1234,repe=5,
  size=c(5),decay=c(0.1),repeticiones=5,itera=200)

   medias2$modelo="Grupo1-avnnet"

# Grupo 2 redes
medias3<-cruzadaavnnetbin(data=em_save_TRAINbis,
```

```
vardep="Churn",listconti=c("OPT_IMP_REP_TotalCharges","TI_OnlineSecuri
ty1","TI_PaymentMethod3","TI_StreamingTV3","TI_MultipleLines1","TI_Con
tract2","TI_StreamingMovies3"),
  listclass=c(""),grupos=4,sinicio=1234,repe=5,
  size=c(7),decay=c(0.01),repeticiones=5,itera=200)

   medias3$modelo="Grupo2-avnnet"

# Grupo 3 redes
medias4<-cruzadaavnnetbin(data=em_save_TRAINbis,

vardep="Churn",listconti=c("OPT_REP_MonthlyCharges","TI_InternetServic
e2","TI_PaperlessBilling1","TI_SeniorCitizen1"),
  listclass=c(""),grupos=4,sinicio=1234,repe=5,
  size=c(5),decay=c(0.01),repeticiones=5,itera=200)

  medias4$modelo="Grupo3-avnnet"

# Grupo 4 redes
medias5<-cruzadaavnnetbin(data=em_save_TRAINbis,

vardep="Churn",listconti=c("LOG_REP_tenure","OPT_IMP_REP_TotalCharges"
,"TI_Contract1","TI_OnlineSecurity1","TI_InternetService2","TI_Paperle
ssBilling1"),
  listclass=c(""),grupos=4,sinicio=1234,repe=5,
  size=c(15),decay=c(0.01),repeticiones=5,itera=200)

  medias5$modelo="Grupo4-avnnet"

# Grupo 5 redes
medias6<-cruzadaavnnetbin(data=em_save_TRAINbis,

vardep="Churn",listconti=c("LOG_REP_tenure","OPT_IMP_REP_TotalCharges"
,"TI_Contract1","TI_TechSupport1","TI_OnlineSecurity1","TI_PaymentMeth
od3","TI_StreamingTV3","TI_MultipleLines1","TI_Contract2","TI_Streamin
gMovies3"),
  listclass=c(""),grupos=4,sinicio=1234,repe=5,
  size=c(7),decay=c(0.01),repeticiones=5,itera=200)

  medias6$modelo="Grupo5-avnnet"

  # grafico de las 5 redes
 union1<-rbind(medias2,medias3,medias4,medias5,medias6)

 par(cex.axis=0.5)
 boxplot(data=union1,tasa~modelo,main="TASA FALLOS")
 boxplot(data=union1,auc~modelo,main="AUC")

 #grafico omitinedo el grupo 3 (tasa de fallos alta)

 union2<-rbind(medias2,medias3,medias5,medias6)

 par(cex.axis=0.5)
 boxplot(data=union2,tasa~modelo,main="TASA FALLOS")
 boxplot(data=union2,auc~modelo,main="AUC")
```

Baggin en R

library(gbm)

```
library(caret)
set.seed(12345)

gbmgrid<-expand.grid(shrinkage=c(0.1,0.05,0.03,0.01,0.001),
 n.minobsinnode=c(5,10,20),
 n.trees=c(100,500,1000,5000),
 interaction.depth=c(2))
control<-trainControl(method = "cv",number=4,savePredictions = "all",
 classProbs=TRUE)
gbm<- train(factor(Churn)~.,data=Set_Grupo_1,
 method="gbm",trControl=control,tuneGrid=gbmgrid,
 distribution="bernoulli", bag.fraction=1,verbose=FALSE)
gbm
 plot(gbm)

# ESTUDIO DE EARLY STOPPING
# Probamos a fijar algunos parámetros para ver como evoluciona
# en función de las iteraciones

gbmgrid<-expand.grid(shrinkage=c(0.03),
 n.minobsinnode=c(5),
 n.trees=c(100,300,500,800,1000,1200,2000,5000),
 interaction.depth=c(2))

control<-trainControl(method = "cv",number=4,savePredictions = "all",
 classProbs=TRUE)

gbm<- train(factor(Churn)~.,data=Set_Grupo_1,
 method="gbm",trControl=control,tuneGrid=gbmgrid,
 distribution="bernoulli", bag.fraction=1,verbose=FALSE)

gbm
plot(gbm,ylim=c(0.75,0.80))

# IMPORTANCIA DE VARIABLES
par(cex=1.3)
summary(gbm)

tabla<-summary(gbm)
par(cex=1.5,las=2)
barplot(tabla$rel.inf,names.arg=row.names(tabla))

# La función cruzadagbmbin permite plantear gradient boosting para binarias

medias17<-cruzadagbmbin(data=Set_Grupo_1, vardep="Churn",
   listconti=c("LOG_REP_tenure", "TI_Contract1", "TI_TechSupport1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=5,shrinkage=0.03,n.trees=300,interaction.depth=2)
medias17$modelo="Grupo 1-gbm"

medias18<-cruzadagbmbin(data=Set_Grupo_2, vardep="Churn",
   listconti=c("OPT_IMP_REP_TotalCharges", "TI_OnlineSecurity1", "TI_PaymentMethod3",
"TI_Contract2"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=5,shrinkage=0.1,n.trees=100,interaction.depth=2)

medias18$modelo="Grupo 2-gbm"
```

```
medias19<-cruzadagbmbin(data=Set_Grupo_3, vardep="Churn",
   listconti=c("TI_InternetService2", "TI_PaperlessBilling1", "OPT_REP_MonthlyCharges"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=5,shrinkage=0.01,n.trees=800,interaction.depth=2)
medias19$modelo="Grupo 3-gbm"

medias20<-cruzadagbmbin(data=Set_Grupo_4, vardep="Churn",
   listconti=c("LOG_REP_tenure", "TI_Contract1", "TI_OnlineSecurity1", "TI_InternetService2"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=20,shrinkage=0.05,n.trees=300,interaction.depth=2)
medias20$modelo="Grupo 4-gbm"

medias21<-cruzadagbmbin(data=Set_Grupo_5, vardep="Churn",
   listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges", "TI_Contract1",
"TI_TechSupport1", "TI_OnlineSecurity1", "TI_PaymentMethod3", "TI_StreamingMovies3"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=10,shrinkage=0.05,n.trees=500,interaction.depth=2)
medias21$modelo="Grupo 5-gbm"
```

Random forest en R

```
library(caret)
library(randomForest)

set.seed(12345)
rfgrid<-expand.grid(mtry=c(2,3))

control<-trainControl(method = "cv",number=4,savePredictions = "all",
 classProbs=TRUE)

rf<- train(factor(Churn)~.,data=Set_Grupo_1,
 method="rf",trControl=control,tuneGrid=rfgrid,
 linout = FALSE,ntree=1000,sampsize=200,nodesize=10,replace=TRUE,
 importance=TRUE)

rf

# IMPORTANCIA DE VARIABLES

final<-rf$finalModel

tabla<-as.data.frame(importance(final))
tabla<-tabla[order(-tabla$MeanDecreaseAccuracy),]
tabla

barplot(tabla$MeanDecreaseAccuracy,names.arg=rownames(tabla))

# PARA PLOTEAR EL ERROR OOB A MEDIDA QUE AVANZAN LAS ITERACIONES
# SE USA DIRECTAMENTE EL PAQUETE randomForest

library(randomForest)
set.seed(12345)

rfbis<-randomForest(factor(Churn)~.,
 data=Set_Grupo_1,
 mtry=2,ntree=5000,sampsize=300,nodesize=10,replace=TRUE)
```

```
plot(rfbis$err.rate[,1])

# TUNEADO BÁSICO DEL TAMAÑO DE MUESTRA A SORTEAR

for (muestra in seq(100,450,50))
{
 # controlamos la semilla pues bagging depende de ella
 set.seed(12345)
rfbis<-randomForest(factor(Churn)~.,
 data=Set_Grupo_1,
 mtry=2,ntree=600,sampsize=muestra,nodesize=10,replace=TRUE)

plot(rfbis$err.rate[,1],main=muestra,ylim=c(0.25,0.5))

}

# Ahora se comprueba con validación cruzada con caret

rfgrid<-expand.grid(mtry=c(2))

rf<- train(factor(Churn)~.,data=Set_Grupo_1,
 method="rf",trControl=control,tuneGrid=rfgrid,
 linout = FALSE,ntree=600,sampsize=450,nodesize=10,replace=TRUE)

rf

# La función cruzadarfbin permite plantear random forest


  medias12<-cruzadarfbin(data=Set_Grupo_1, vardep="Churn",
  listconti=c("LOG_REP_tenure", "TI_Contract1","TI_TechSupport1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,nodesize=10,
  mtry=2,ntree=600,replace=TRUE)
  medias12$modelo="Grupo 1-rf"

  medias13<-cruzadarfbin(data=Set_Grupo_2, vardep="Churn",
listconti=c("OPT_IMP_REP_TotalCharges","TI_OnlineSecurity1","TI_PaymentMethod3","TI_Streamin
gTV3","TI_Contract2","TI_StreamingMovies3"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,nodesize=10,
  mtry=6,ntree=1000,replace=TRUE)
  medias13$modelo="Grupo 2-rf"

  medias14<-cruzadarfbin(data=Set_Grupo_3, vardep="Churn",
  listconti=c("TI_InternetService2", "TI_PaperlessBilling1", "TI_SeniorCitizen1",
"OPT_REP_MonthlyCharges"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,nodesize=10,
  mtry=3,ntree=3000,replace=TRUE)
 medias14$modelo="Grupo 3-rf"

  medias15<-cruzadarfbin(data=Set_Grupo_4, vardep="Churn",
  listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges", "TI_Contract1",
"TI_OnlineSecurity1", "TI_InternetService2", "TI_PaperlessBilling1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,nodesize=10,
  mtry=5,ntree=800,replace=TRUE)
medias15$modelo="Grupo 4-rf
```

```
   medias16<-cruzadarfbin(data=Set_Grupo_5, vardep="Churn",
  listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges","TI_Contract1", "TI_TechSupport1",
"TI_OnlineSecurity1", "TI_PaymentMethod3", "TI_StreamingTV3", "TI_MultipleLines1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,nodesize=10,
  mtry=9,ntree=2500,replace=TRUE)
 medias16$modelo="Grupo 5-rf"
```

Gradient boosting en R

```
library(gbm)
library(caret)
set.seed(12345)

gbmgrid<-expand.grid(shrinkage=c(0.1,0.05,0.03,0.01,0.001),
 n.minobsinnode=c(5,10,20),
 n.trees=c(100,500,1000,5000),
 interaction.depth=c(2))

control<-trainControl(method = "cv",number=4,savePredictions = "all",
 classProbs=TRUE)

gbm<- train(factor(Churn)~.,data=Set_Grupo_1,
 method="gbm",trControl=control,tuneGrid=gbmgrid,
 distribution="bernoulli", bag.fraction=1,verbose=FALSE)

gbm

plot(gbm)

# ESTUDIO DE EARLY STOPPING

gbmgrid<-expand.grid(shrinkage=c(0.03),
 n.minobsinnode=c(5),
 n.trees=c(100,300,500,800,1000,1200,2000,5000),
 interaction.depth=c(2))

control<-trainControl(method = "cv",number=4,savePredictions = "all",
 classProbs=TRUE)

gbm<- train(factor(Churn)~.,data=Set_Grupo_1,
 method="gbm",trControl=control,tuneGrid=gbmgrid,
 distribution="bernoulli", bag.fraction=1,verbose=FALSE)

gbm
plot(gbm,ylim=c(0.75,0.80))

# IMPORTANCIA DE VARIABLES
par(cex=1.3)
summary(gbm)

tabla<-summary(gbm)
par(cex=1.5,las=2)
barplot(tabla$rel.inf,names.arg=row.names(tabla))
```

```
# La función cruzadagbmbin permite plantear gradient boosting para binarias

medias17<-cruzadagbmbin(data=Set_Grupo_1, vardep="Churn",
   listconti=c("LOG_REP_tenure", "TI_Contract1", "TI_TechSupport1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=5,shrinkage=0.03,n.trees=300,interaction.depth=2)
medias17$modelo="Grupo 1-gbm"

medias18<-cruzadagbmbin(data=Set_Grupo_2, vardep="Churn",
   listconti=c("OPT_IMP_REP_TotalCharges", "TI_OnlineSecurity1", "TI_PaymentMethod3",
"TI_Contract2"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=5,shrinkage=0.1,n.trees=100,interaction.depth=2)
medias18$modelo="Grupo 2-gbm"

medias19<-cruzadagbmbin(data=Set_Grupo_3, vardep="Churn",
   listconti=c("TI_InternetService2", "TI_PaperlessBilling1", "OPT_REP_MonthlyCharges"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=5,shrinkage=0.01,n.trees=800,interaction.depth=2)
medias19$modelo="Grupo 3-gbm"

medias20<-cruzadagbmbin(data=Set_Grupo_4, vardep="Churn",
   listconti=c("LOG_REP_tenure", "TI_Contract1", "TI_OnlineSecurity1", "TI_InternetService2"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=20,shrinkage=0.05,n.trees=300,interaction.depth=2)
medias20$modelo="Grupo 4-gbm"

medias21<-cruzadagbmbin(data=Set_Grupo_5, vardep="Churn",
   listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges", "TI_Contract1",
"TI_TechSupport1", "TI_OnlineSecurity1", "TI_PaymentMethod3", "TI_StreamingMovies3"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
n.minobsinnode=10,shrinkage=0.05,n.trees=500,interaction.depth=2)
medias21$modelo="Grupo 5-gbm"
```

SVM en R

```
# TUNEADO SVM BINARIA
# ***************************
install.packages("kernlab")
library(kernlab)
library(caret)


#  SVM LINEAL: SOLO PARÁMETRO C

set.seed(12345)
SVMgrid<-expand.grid(C=c(0.01,0.05,0.1,0.2,0.5,1,2,5,10))

control<-trainControl(method = "cv",number=4,savePredictions = "all")

SVM<- train(data=Set_Grupo_1,factor(Churn)~LOG_REP_tenure+TI_Contract1+TI_TechSupport1,
  method="svmLinear",trControl=control,
 tuneGrid=SVMgrid,verbose=FALSE)
```

```
SVM$results
plot(SVM$results$C,SVM$results$Accuracy)

# Rehago el grid para observar mejor el intervalo de C entre 0 y 0.6
SVMgrid<-expand.grid(C=c(0.01,0.02,0.03,0.04,0.05,0.8,0.1,0.2,0.3,0.4,0.5,0.6))

control<-trainControl(method = "cv",number=4,savePredictions = "all")

SVM<- train(data=Set_Grupo_1,factor(Churn)~LOG_REP_tenure+TI_Contract1+TI_TechSupport1,
  method="svmLinear",trControl=control,
 tuneGrid=SVMgrid,verbose=FALSE)

SVM$results
plot(SVM$results$C,SVM$results$Accuracy)


#  SVM Polinomial: PARÁMETROS C, degree, scale

SVMgrid<-expand.grid(C=c(0.01,0.05,0.1,0.2,0.5,1,2,5,10),
 degree=c(2,3),scale=c(0.1,0.5,1,2,5))

control<-trainControl(method = "cv",
 number=4,savePredictions = "all")


SVM<- train(data=Set_Grupo_1,factor(Churn)~LOG_REP_tenure+TI_Contract1+TI_TechSupport1,
  method="svmPoly",trControl=control,
 tuneGrid=SVMgrid,verbose=FALSE)

SVM

SVM$results

# LOS GRÁFICOS DOS A DOS NO SIRVEN
# plot(SVM$results$C,SVM$results$Accuracy)
# plot(SVM$results$degree,SVM$results$Accuracy)
# plot(SVM$results$scale,SVM$results$Accuracy)

dat<-as.data.frame(SVM$results)
library(ggplot2)

# PLOT DE DOS VARIABLES CATEGÓRICAS, UNA CONTINUA
ggplot(dat, aes(x=factor(C), y=Accuracy,
 color=factor(degree),pch=factor(scale))) +
  geom_point(position=position_dodge(width=0.5),size=3)

# SOLO DEGREE=2
dat2<-dat[dat$degree==2,]

ggplot(dat2, aes(x=factor(C), y=Accuracy,
 colour=factor(scale))) +
  geom_point(position=position_dodge(width=0.5),size=3)

#  SVM RBF: PARÁMETROS C, sigma

SVMgrid<-expand.grid(C=c(0.01,0.05,0.1,0.2,0.5,1,2,5,10,30),
 sigma=c(0.01,0.05,0.1,0.2,0.5,1,2,5,10,30))

control<-trainControl(method = "cv",
```

```
 number=4,savePredictions = "all")


SVM<- train(data=Set_Grupo_1,factor(Churn)~LOG_REP_tenure+TI_Contract1+TI_TechSupport1,
  method="svmRadial",trControl=control,
 tuneGrid=SVMgrid,verbose=FALSE)

SVM

dat<-as.data.frame(SVM$results)

ggplot(dat, aes(x=factor(C), y=Accuracy,
 color=factor(sigma)))+
 geom_point(position=position_dodge(width=0.5),size=3)


# cruzada para la generacion de graficas consolidado de grupos

medias22<-cruzadaSVMbin(data=Set_Grupo_1, vardep="Churn",
   listconti=c("LOG_REP_tenure", "TI_Contract1", "TI_TechSupport1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,C=0.01)
medias22$modelo="Grupo 1-SVM"

medias23<-cruzadaSVMbinPoly(data=Set_Grupo_1, vardep="Churn",
   listconti=c("LOG_REP_tenure", "TI_Contract1", "TI_TechSupport1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=0.01,degree=2,scale=1)
medias23$modelo=" Grupo 1-SVMPoly"

medias24<-cruzadaSVMbinRBF(data=Set_Grupo_1, vardep="Churn",
   listconti=c("LOG_REP_tenure", "TI_Contract1", "TI_TechSupport1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=0.5,sigma=0.2)
medias24$modelo="Grupo 1 SVMRBF"

#modelos de grupo 2

medias25<-cruzadaSVMbin(data=Set_Grupo_2, vardep="Churn",
   listconti=c("OPT_IMP_REP_TotalCharges", "TI_OnlineSecurity1", "TI_PaymentMethod3",
"TI_Contract2", "TI_StreamingTV3", "TI_StreamingMovies3", "TI_MultipleLines1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,C=0.02)
medias25$modelo="Grupo 2-SVM"

medias26<-cruzadaSVMbinPoly(data=Set_Grupo_2, vardep="Churn",
   listconti=c("OPT_IMP_REP_TotalCharges", "TI_OnlineSecurity1", "TI_PaymentMethod3",
"TI_Contract2", "TI_StreamingTV3", "TI_StreamingMovies3", "TI_MultipleLines1"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=0.2,degree=3,scale=0.1)
medias26$modelo=" Grupo 2-SVMPoly"

medias27<-cruzadaSVMbinRBF(data=Set_Grupo_2, vardep="Churn",
  listconti=c("OPT_IMP_REP_TotalCharges", "TI_OnlineSecurity1", "TI_PaymentMethod3",
"TI_Contract2", "TI_StreamingTV3", "TI_StreamingMovies3", "TI_MultipleLines1"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,C=1,sigma=2)
```

```
medias27$modelo="Grupo 2 SVMRBF"

#modelos de grupo 3

medias28<-cruzadaSVMbin(data=Set_Grupo_3, vardep="Churn",
   listconti=c("TI_InternetService2", "TI_PaperlessBilling1", "TI_SeniorCitizen1",
"OPT_REP_MonthlyCharges"),
 listclass=c(""),
  grupos=4,sinicio=1234,repe=5,C=0.01)
medias28$modelo="Grupo 3-SVM"

medias29<-cruzadaSVMbinPoly(data=Set_Grupo_3, vardep="Churn",
  listconti=c("TI_InternetService2", "TI_PaperlessBilling1", "TI_SeniorCitizen1",
"OPT_REP_MonthlyCharges"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=0.5,degree=2,scale=2)
medias29$modelo="Grupo 3-SVMPoly"

medias30<-cruzadaSVMbinRBF(data=Set_Grupo_3, vardep="Churn",
  listconti=c("TI_InternetService2", "TI_PaperlessBilling1",
"TI_SeniorCitizen1","OPT_REP_MonthlyCharges"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=0.5,sigma=0.5)
medias30$modelo="Grupo 3 -SVMRBF"


#modelos de grupo 4


medias31<-cruzadaSVMbin(data=Set_Grupo_4, vardep="Churn",
  listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges", "TI_Contract1",
"TI_OnlineSecurity1", "TI_InternetService2", "TI_PaperlessBilling1"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,C=0.03)
medias31$modelo="Grupo 4-SVM"


medias32<-cruzadaSVMbinPoly(data=Set_Grupo_4, vardep="Churn",
  listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges", "TI_Contract1",
"TI_OnlineSecurity1", "TI_InternetService2", "TI_PaperlessBilling1"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=0.2,degree=3,scale=1)
medias32$modelo="Grupo 4-SVMPoly"


medias33<-cruzadaSVMbinRBF(data=Set_Grupo_4, vardep="Churn",
  listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges", "TI_Contract1",
"TI_OnlineSecurity1", "TI_InternetService2", "TI_PaperlessBilling1"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=0.5,sigma=0.5)
medias33$modelo="Grupo 4 -SVMRBF"


#modelos de grupo 5
```

```
medias34<-cruzadaSVMbin(data=Set_Grupo_5, vardep="Churn",
  listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges","TI_Contract1", "TI_TechSupport1",
"TI_OnlineSecurity1", "TI_PaymentMethod3", "TI_StreamingTV3", "TI_MultipleLines1",
"TI_Contract2", "TI_StreamingMovies3"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,C=0.2)
medias34$modelo="Grupo 5-SVM"


medias35<-cruzadaSVMbinPoly(data=Set_Grupo_5, vardep="Churn",
  listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges","TI_Contract1", "TI_TechSupport1",
"TI_OnlineSecurity1", "TI_PaymentMethod3", "TI_StreamingTV3", "TI_MultipleLines1",
"TI_Contract2", "TI_StreamingMovies3"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=0.5,degree=3,scale=0.5)
medias35$modelo=" Grupo 5-SVMPoly"


medias36<-cruzadaSVMbinRBF(data=Set_Grupo_5, vardep="Churn",
  listconti=c("LOG_REP_tenure", "OPT_IMP_REP_TotalCharges","TI_Contract1", "TI_TechSupport1",
"TI_OnlineSecurity1", "TI_PaymentMethod3", "TI_StreamingTV3", "TI_MultipleLines1",
"TI_Contract2", "TI_StreamingMovies3"),
  listclass=c(""),
  grupos=4,sinicio=1234,repe=5,
  C=5,sigma=0.01)
medias36$modelo="Grupo 5 SVMRBF"


#grafico de cajas de modelos


union8<-
rbind(medias22,medias23,medias24,medias25,medias26,medias27,medias28,medias29,medias30,medias3
1,medias32,medias33,medias34,medias35,medias36)

par(cex.axis=0.8)
boxplot(data=union8,tasa~modelo,main="TASA FALLOS",col="pink")
boxplot(data=union8,auc~modelo,main="AUC",col="pink")

uni<-union8
uni$modelo <- with(uni,
 reorder(modelo,tasa, median))
par(cex.axis=0.5,las=2)
boxplot(data=uni,tasa~modelo,col="pink",main="TASA FALLOS")

uni<-union8
uni$modelo <- with(uni,
 reorder(modelo,auc, median))
par(cex.axis=0.5,las=2)
boxplot(data=uni,auc~modelo,col="pink",main="AUC")


#grafico de cajas de modelos grupo 4 y 5 los mejores

union9<-rbind(medias31,medias32,medias33,medias34,medias35,medias36)

uni<-union9
uni$modelo <- with(uni,
 reorder(modelo,tasa, median))
```

```
par(cex.axis=0.5,las=2)
boxplot(data=uni,tasa~modelo,col="pink",main="TASA FALLOS")
```

Ensamblado en R

```
# APLICACIÓN CRUZADAS PARA ENSAMBLAR

medias1<-cruzadalogistica(data=archivo,
 vardep=vardep,listconti=listconti,
 listclass=listclass,grupos=grupos,sinicio=sinicio,repe=repe)

medias1bis<-as.data.frame(medias1[1])
medias1bis$modelo<-"Logistica"
predi1<-as.data.frame(medias1[2])
predi1$logi<-predi1$Yes

medias2<-cruzadaavnnetbin(data=archivo,
 vardep=vardep,listconti=listconti,
 listclass=listclass,grupos=grupos,sinicio=sinicio,repe=repe,
    size=c(5),decay=c(0.1),repeticiones=5,itera=200)

medias2bis<-as.data.frame(medias2[1])
medias2bis$modelo<-"avnnet"
predi2<-as.data.frame(medias2[2])
predi2$avnnet<-predi2$Yes


medias3<-cruzadarfbin(data=archivo,
  vardep=vardep,listconti=listconti,
  listclass=listclass,grupos=grupos,sinicio=sinicio,repe=repe,
 mtry=3,ntree=200,nodesize=10,replace=TRUE)


medias3bis<-as.data.frame(medias3[1])
medias3bis$modelo<-"rf"
predi3<-as.data.frame(medias3[2])
predi3$rf<-predi3$Yes

medias4<-cruzadagbmbin(data=archivo,
  vardep=vardep,listconti=listconti,
  listclass=listclass,grupos=grupos,sinicio=sinicio,repe=repe,
n.minobsinnode=5,shrinkage=0.001,n.trees=3000,interaction.depth=2)

medias4bis<-as.data.frame(medias4[1])
medias4bis$modelo<-"gbm"
predi4<-as.data.frame(medias4[2])
predi4$gbm<-predi4$Yes

medias5<-cruzadaxgbmbin(data=archivo,
  vardep=vardep,listconti=listconti,
  listclass=listclass,grupos=grupos,sinicio=sinicio,repe=repe,
   min_child_weight=10,eta=0.08,nrounds=100,max_depth=6,
  gamma=0,colsample_bytree=1,subsample=1,
 alpha=0,lambda=0,lambda_bias=0)


medias5bis<-as.data.frame(medias5[1])
medias5bis$modelo<-"xgbm"
predi5<-as.data.frame(medias5[2])
```

```
predi5$xgbm<-predi5$Yes


medias6<-cruzadaSVMbin(data=archivo,
   vardep=vardep,listconti=listconti,
   listclass=listclass,grupos=grupos,
 sinicio=sinicio,repe=repe,C=0.03)

medias6bis<-as.data.frame(medias6[1])
medias6bis$modelo<-"svmLinear"
predi6<-as.data.frame(medias6[2])
predi6$svmLinear<-predi6$Yes


medias7<-cruzadaSVMbinPoly(data=archivo,
   vardep=vardep,listconti=listconti,
   listclass=listclass,grupos=grupos,sinicio=sinicio,repe=repe,
 C=0.02,degree=2,scale=2)

medias7bis<-as.data.frame(medias7[1])
medias7bis$modelo<-"svmPoly"
predi7<-as.data.frame(medias7[2])
predi7$svmPoly<-predi7$Yes

medias8<-cruzadaSVMbinRBF(data=archivo,
   vardep=vardep,listconti=listconti,
   listclass=listclass,grupos=grupos,
 sinicio=sinicio,repe=repe,
 C=5,sigma=0.01)

medias8bis<-as.data.frame(medias8[1])
medias8bis$modelo<-"svmRadial"
predi8<-as.data.frame(medias8[2])
predi8$svmRadial<-predi8$Yes

union1<-rbind(medias1bis,medias2bis,
 medias3bis,medias4bis,medias5bis,medias6bis,
 medias7bis,medias8bis)

par(cex.axis=0.8)
boxplot(data=union1,tasa~modelo,col="pink",main='TASA FALLOS')
boxplot(data=union1,auc~modelo,col="pink",main='AUC')

# CONSTRUCCIÓN DE TODOS LOS ENSAMBLADOS
# SE UTILIZARÁN LOS ARCHIVOS SURGIDOS DE LAS FUNCIONES LLAMADOS predi1,...

unipredi<-cbind(predi1,predi2,predi3,predi4,predi5,predi6,predi7,predi8)

# Esto es para eliminar columnas duplicadas
unipredi<- unipredi[, !duplicated(colnames(unipredi))]

# Construccion de ensamblados, cambiar al gusto

unipredi$predi9<-(unipredi$logi+unipredi$avnnet)/2
unipredi$predi10<-(unipredi$logi+unipredi$rf)/2
unipredi$predi11<-(unipredi$logi+unipredi$gbm)/2
unipredi$predi12<-(unipredi$logi+unipredi$xgbm)/2
unipredi$predi13<-(unipredi$logi+unipredi$svmLinear)/2
unipredi$predi14<-(unipredi$logi+unipredi$svmPoly)/2
unipredi$predi15<-(unipredi$logi+unipredi$svmRadial)/2
```

```
unipredi$predi16<-(unipredi$avnnet+unipredi$rf)/2
unipredi$predi17<-(unipredi$avnnet+unipredi$gbm)/2
unipredi$predi18<-(unipredi$avnnet+unipredi$xgbm)/2
unipredi$predi19<-(unipredi$avnnet+unipredi$svmLinear)/2
unipredi$predi20<-(unipredi$avnnet+unipredi$svmPoly)/2
unipredi$predi21<-(unipredi$avnnet+unipredi$svmRadial)/2
unipredi$predi22<-(unipredi$rf+unipredi$gbm)/2
unipredi$predi23<-(unipredi$rf+unipredi$xgbm)/2
unipredi$predi24<-(unipredi$rf+unipredi$svmLinear)/2
unipredi$predi25<-(unipredi$rf+unipredi$svmPoly)/2
unipredi$predi26<-(unipredi$rf+unipredi$svmRadial)/2
unipredi$predi27<-(unipredi$gbm+unipredi$xgbm)/2
unipredi$predi28<-(unipredi$gbm+unipredi$svmLinear)/2
unipredi$predi29<-(unipredi$gbm+unipredi$svmPoly)/2
unipredi$predi30<-(unipredi$gbm+unipredi$svmRadial)/2

unipredi$predi31<-(unipredi$logi+unipredi$avnnet+unipredi$rf)/3
unipredi$predi32<-(unipredi$logi+unipredi$avnnet+unipredi$gbm)/3
unipredi$predi33<-(unipredi$logi+unipredi$avnnet+unipredi$xgbm)/3
unipredi$predi34<-(unipredi$logi+unipredi$avnnet+unipredi$svmLinear)/3
unipredi$predi35<-(unipredi$logi+unipredi$avnnet+unipredi$svmPoly)/3
unipredi$predi36<-(unipredi$logi+unipredi$avnnet+unipredi$svmRadial)/3
unipredi$predi37<-(unipredi$logi+unipredi$rf+unipredi$gbm)/3
unipredi$predi38<-(unipredi$logi+unipredi$rf+unipredi$xgbm)/3
unipredi$predi39<-(unipredi$logi+unipredi$rf+unipredi$svmLinear)/3
unipredi$predi40<-(unipredi$logi+unipredi$rf+unipredi$svmPoly)/3
unipredi$predi41<-(unipredi$logi+unipredi$rf+unipredi$svmRadial)/3
unipredi$predi42<-(unipredi$logi+unipredi$gbm+unipredi$xgbm)/3
unipredi$predi43<-(unipredi$logi+unipredi$gbm+unipredi$xgbm)/3
unipredi$predi44<-(unipredi$logi+unipredi$gbm+unipredi$svmLinear)/3
unipredi$predi45<-(unipredi$logi+unipredi$gbm+unipredi$svmPoly)/3
unipredi$predi46<-(unipredi$logi+unipredi$gbm+unipredi$svmRadial)/3
unipredi$predi47<-(unipredi$logi+unipredi$xgbm+unipredi$svmLinear)/3
unipredi$predi48<-(unipredi$logi+unipredi$xgbm+unipredi$svmPoly)/3
unipredi$predi49<-(unipredi$logi+unipredi$xgbm+unipredi$svmRadial)/3

unipredi$predi50<-(unipredi$rf+unipredi$gbm+unipredi$svmLinear)/3
unipredi$predi51<-(unipredi$rf+unipredi$gbm+unipredi$svmPoly)/3
unipredi$predi52<-(unipredi$rf+unipredi$gbm+unipredi$svmRadial)/3

unipredi$predi53<-(unipredi$rf+unipredi$xgbm+unipredi$svmLinear)/3
unipredi$predi54<-(unipredi$rf+unipredi$xgbm+unipredi$svmPoly)/3
unipredi$predi55<-(unipredi$rf+unipredi$xgbm+unipredi$svmRadial)/3

unipredi$predi56<-(unipredi$rf+unipredi$avnnet+unipredi$gbm)/3
unipredi$predi57<-(unipredi$rf+unipredi$avnnet+unipredi$xgbm)/3
unipredi$predi58<-(unipredi$rf+unipredi$avnnet+unipredi$svmLinear)/3
unipredi$predi59<-(unipredi$rf+unipredi$avnnet+unipredi$svmPoly)/3
unipredi$predi60<-(unipredi$rf+unipredi$avnnet+unipredi$svmRadial)/3

unipredi$predi61<-(unipredi$avnnet+unipredi$gbm+unipredi$svmLinear)/3
unipredi$predi62<-(unipredi$avnnet+unipredi$gbm+unipredi$svmPoly)/3
unipredi$predi63<-(unipredi$avnnet+unipredi$gbm+unipredi$svmRadial)/3

unipredi$predi64<-(unipredi$logi+unipredi$rf+unipredi$gbm+unipredi$avnnet)/4
unipredi$predi65<-(unipredi$logi+unipredi$rf+unipredi$xgbm+unipredi$avnnet)/4
unipredi$predi66<-(unipredi$logi+unipredi$rf+unipredi$xgbm+unipredi$avnnet)/4

unipredi$predi67<-(unipredi$logi+unipredi$rf+unipredi$xgbm+unipredi$avnnet+unipredi$svmLinear)/5
unipredi$predi68<-(unipredi$logi+unipredi$rf+unipredi$xgbm+unipredi$avnnet+unipredi$svmPoly)/5
```

```
unipredi$predi69<-(unipredi$logi+unipredi$rf+unipredi$xgbm+unipredi$avnnet+unipredi$svmRadial)/5


# Listado de modelos a considerar, cambiar al gusto

dput(names(unipredi))

listado<-c("logi", "avnnet",
"rf","gbm", "svmLinear",  "svmPoly",
"svmRadial","predi9", "predi10", "predi11", "predi12",
"predi13", "predi14", "predi15", "predi16", "predi17", "predi18",
"predi19", "predi20", "predi21", "predi22", "predi23", "predi24",
"predi25", "predi26", "predi27", "predi28", "predi29", "predi30",
"predi31", "predi32", "predi33", "predi34", "predi35", "predi36",
"predi37", "predi38", "predi39", "predi40", "predi41", "predi42",
"predi43", "predi44", "predi45", "predi46", "predi47", "predi48",
"predi49", "predi50", "predi51", "predi52", "predi53", "predi54",
"predi55", "predi56", "predi57", "predi58", "predi59", "predi60",
"predi61", "predi62", "predi63", "predi64", "predi65", "predi66",
"predi67", "predi68", "predi69")

# Cambio a Yes, No, todas las predicciones

# Defino funcion tasafallos

tasafallos<-function(x,y) {
  confu<-confusionMatrix(x,y)
  tasa<-confu[[3]][1]
  return(tasa)
}

auc<-function(x,y) {
  curvaroc<-roc(response=x,predictor=y)
  auc<-curvaroc$auc
  return(auc)
 }

# Se obtiene el numero de repeticiones CV y se calculan las medias por repe en
# el data frame medias0

repeticiones<-nlevels(factor(unipredi$Rep))
unipredi$Rep<-as.factor(unipredi$Rep)
unipredi$Rep<-as.numeric(unipredi$Rep)


medias0<-data.frame(c())
for (prediccion in listado)
{
 unipredi$proba<-unipredi[,prediccion]
 unipredi[,prediccion]<-ifelse(unipredi[,prediccion]>0.5,"Yes","No")
 for (repe in 1:repeticiones)
{
 paso <- unipredi[(unipredi$Rep==repe),]
 pre<-factor(paso[,prediccion])
 archi<-paso[,c("proba","obs")]
 archi<-archi[order(archi$proba),]
 obs<-paso[,c("obs")]
 tasa=1-tasafallos(pre,obs)
 t<-as.data.frame(tasa)
 t$modelo<-prediccion
```

```
 auc<-auc(archi$obs,archi$proba)
 t$auc<-auc
 medias0<-rbind(medias0,t)
}
}

# Finalmente boxplot

par(cex.axis=0.5,las=2)
boxplot(data=medias0,tasa~modelo,col="pink",main="TASA FALLOS")

# Para AUC se utiliza la variable auc del archivo medias0

boxplot(data=medias0,auc~modelo,col="pink",main="AUC")

# PRESENTACION TABLA MEDIAS

tablamedias<-medias0 %>%
  group_by(modelo) %>%
  summarize(tasa=mean(tasa))

tablamedias<-tablamedias[order(tablamedias$tasa),]


# ORDENACIÓN DEL FACTOR MODELO POR LAS MEDIAS EN TASA
# PARA EL GRAFICO

medias0$modelo <- with(medias0,
 reorder(modelo,tasa, mean))
par(cex.axis=0.7,las=2)
boxplot(data=medias0,tasa~modelo,col="pink", main='TASA FALLOS')

# ************************************
# PARA AUC
# ************************************

# PRESENTACION TABLA MEDIAS

tablamedias2<-medias0 %>%
  group_by(modelo) %>%
  summarize(auc=mean(auc))

tablamedias2<-tablamedias2[order(-tablamedias2$auc),]


# ORDENACIÓN DEL FACTOR MODELO POR LAS MEDIAS EN AUC
# PARA EL GRAFICO

medias0$modelo <- with(medias0,
 reorder(modelo,auc, mean))
par(cex.axis=0.7,las=2)
boxplot(data=medias0,auc~modelo,col="pink", main='AUC')

# Se pueden escoger listas pero el factor hay que pasarlo a character
# para que no salgan en el boxplot todos los niveles del factor

listadobis<-c("logi", "avnnet",
"rf","gbm",  "xgbm", "svmLinear",  "svmPoly",
"svmRadial","predi45", "predi14", "predi46", "predi47")
```

```
medias0$modelo<-as.character(medias0$modelo)

mediasver<-medias0[medias0$modelo %in% listadobis,]


mediasver$modelo <- with(mediasver,
 reorder(modelo,auc, median))

par(cex.axis=0.9,las=2)
boxplot(data=mediasver,auc~modelo,col="pink",main='AUC')



# GRÁFICOS DE APOYO PARA OBSERVAR COMPORTAMIENTO DE LOS MODELOS

unipredi<-cbind(predi1,predi2,predi3,predi4,predi5,predi6,predi7,predi8)
# Esto es para eliminar columnas duplicadas
unipredi<- unipredi[, !duplicated(colnames(unipredi))]
# Añadir ensamblados
unipredi$predi47<-(unipredi$logi+unipredi$xgbm+unipredi$svmLinear)/3
unipredi$predi14<-(unipredi$logi+unipredi$svmPoly)/2
unipredi$predi45<-(unipredi$logi+unipredi$gbm+unipredi$svmPoly)/3
unipredi$predi46<-(unipredi$logi+unipredi$gbm+unipredi$svmRadial)/3

# Me quedo con la primera repetición de validación cruzada para los análisis
unigraf<-unipredi[unipredi$Rep=="Rep1",]
# Correlaciones entre predicciones de cada algoritmo individual
solos<-c("logi", "avnnet",
"rf","gbm",  "xgbm", "svmLinear",  "svmPoly",
"svmRadial")
mat<-unigraf[,solos]
matrizcorr<-cor(mat)
matrizcorr
library(corrplot)
corrplot(matrizcorr, type = "upper", order = "hclust",
        tl.col = "black", tl.srt = 45,cl.lim=c(0.7,1),is.corr=FALSE)

library(ggplot2)

qplot(svmRadial,logi,data=unigraf,colour=obs)+
 geom_hline(yintercept=0.5, color="black", size=1)+
 geom_vline(xintercept=0.5, color="black", size=1)

qplot(predi47,logi,data=unigraf,colour=obs)+
 geom_hline(yintercept=0.5, color="black", size=1)+
 geom_vline(xintercept=0.5, color="black", size=1)

qplot(gbm,svmPoly,data=unigraf,colour=obs)+
 geom_hline(yintercept=0.5, color="black", size=1)+
 geom_vline(xintercept=0.5, color="black", size=1)
```

#ENSAMBLADO SAS Y R

```
par(cex.axis=0.5,las=2)
boxplot(data=mix_ensamblados,tasa~modelo,main="TASA FALLOS")

mix_ensamblados$modelo <- with(mix_ensamblados, reorder(modelo,tasa, median))

par(cex.axis=0.7,las=2)
```

```
boxplot(data=mix_ensamblados,tasa~modelo,col="pink",main='TASA FALLOS')


listadobis<-c("logi-r", "avnnet-r", "rf-r","gbm-r", "svmLinear-r",  "svmPoly-r", "svmRadial-r","1 LOG",
"2 RED", "4 BOOST")

mix_ensamblados$modelo<-as.character(mix_ensamblados$modelo)
mediasver<-mix_ensamblados[mix_ensamblados$modelo %in% listadobis,]

mediasver$modelo <- with(mediasver,reorder(modelo,tasa, median))
par(cex.axis=0.9,las=2)
boxplot(data=mediasver,tasa~modelo,col="pink",main='TASA FALLOS')
```

Ensanblado en SAS

```
libname ensamble 'C:\Users\michelle\Documents\TFM CHURN\disco
c\ensambado sas\em_save_TRAIN.xlsx';
data uno; set ensamble.em_save_TRAIN;

data uno;set ensamble.em_save_TRAIN;Churn2=Churn*1;drop Churn;run;
data uno;Set uno;Churn=Churn2;drop Churn2;run;

%cruzadastackcon(archivo=uno,
vardepen=Churn,
listconti=TI_Partner1 TI_PaperlessBilling2 TI_Partner2
TI_OnlineSecurity3 TI_PaperlessBilling1
        TI_PaymentMethod3 TI_PaymentMethod4 TI_PaymentMethod1
TI_PaymentMethod2 TI_MultipleLines3
        TI_OnlineBackup1 TI_MultipleLines1 TI_MultipleLines2
TI_OnlineSecurity1 TI_OnlineSecurity2
        TI_OnlineBackup2 TI_OnlineBackup3 TI_TechSupport1
TI_TechSupport2 TI_StreamingTV2
        TI_StreamingTV3 TI_gender2 numMissing TI_TechSupport3
TI_gender1 TI_SeniorCitizen1
        TI_SeniorCitizen2 TI_PhoneService1 TI_PhoneService2
TI_StreamingMovies3 TI_StreamingTV1
        TI_StreamingMovies1 TI_StreamingMovies2 TI_Contract2
TI_Contract1 TI_Dependents1 TI_Contract3
        LOG_REP_tenure EXP_var_aleatoria OPT_IMP_REP_TotalCharges
OPT_REP_MonthlyCharges
        TI_InternetService1 TI_DeviceProtection3 TI_InternetService3
TI_InternetService2
        TI_DeviceProtection1 TI_Dependents2 TI_DeviceProtection2
OPT_IMP_REP_TotalCharges OPT_REP_MonthlyCharges,
ngrupos=5,seminicio=22345,semifinal=22349,
nodos=7,algo=levmar mom=0.2 learn=0.7,rediter=100,/*parametros red */
maxtrees=100,vars_to_try=8,trainfraction=0.7,leafsize=3,maxdepth=4,/*r
andom forest */
bleafsize=15,iterations=300,bmaxbranch=4,bmaxdepth=4,shrinkage=0.05,/*
g boosting*/
kernel=lineal,degree=2,k_par=1,c=10);

/* CORRELACIONES ENTRE PREDICCIONES PUNTUALES ULTIMA SEMILLA Y GRUPO*/
proc corr data=salfin;var predi1-predi5;run;



/* EJEMPLO CON SVM */
```

```
/*PREPARACION GRAFICO Y ETIQUETAS */

data cajas;
array ase{23};
set final;
do i=1 to 23;
modelo=i;
error=ase{i};
output;
end;
run;

/* EN ESTAS OPCIONES SE CAMBIA LA LETRA Y LA ALTURA DEL TEXTO EN LOS
EJES CON HTEXT.
options font="Courier New" bold 8;
run;goptions htext=8pt;
*/

proc sort data=cajas;by modelo;
data eti;length eti $ 13;
input modelo eti $;
cards;
1 LOG
2 RED
3 RFOR
4 BOOST
5 SVM
6 RLOG
7 REDFOR
8 REDBOO
9 LRFOR
10 LBOOST
11 RFORBOO
12 R-L-RFOR
13 R-L-BOO
14 R-RF-BOO
15 L-RF-BOO
16 R-L-RF-BOO
17 15ponde
18 R-SVM
19 RF-SVM
20 L-SVM
21 BOO-SVM
22 SVMLRF
23 RLRFBSVM
;
data cajas2;merge cajas eti;by modelo;
title1
h=2 box=1 j=c c=red 'Prediccion Churn' j=c ;

options font="Courier New" bold 8;
run;goptions htext=5pt;

ods graphics off;

proc boxplot data=cajas2;plot error*ETI /
cboxes        = dagr
cboxfill      = ywh;
/* vaxis=0.20 to 0.35 by 0.01 */
;run;
```